Misuse and Misapplication in Statistical Data Analysis

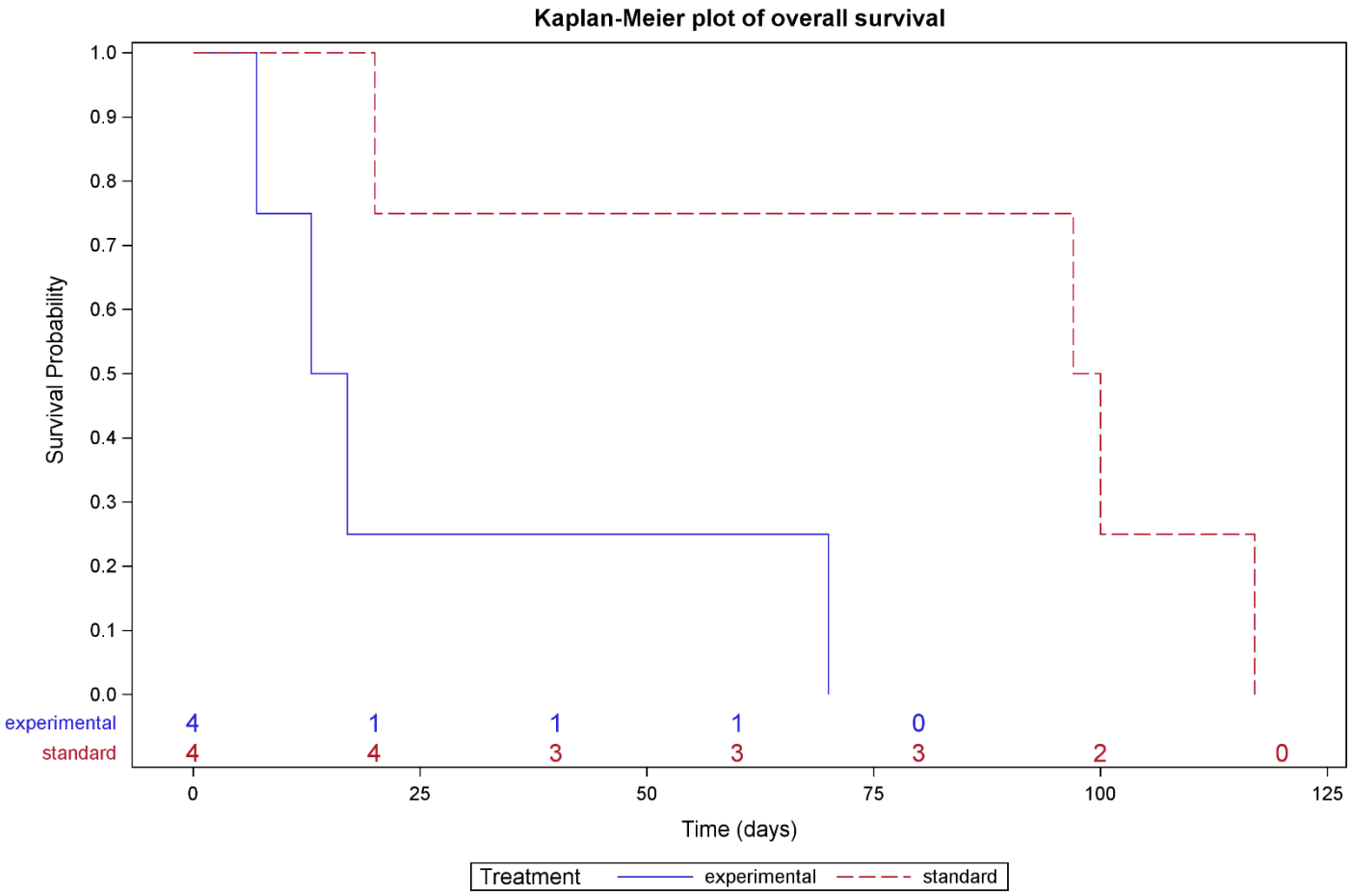
– A topic that never goes out of style

Corinna Miede

PhUSE SDE Basel

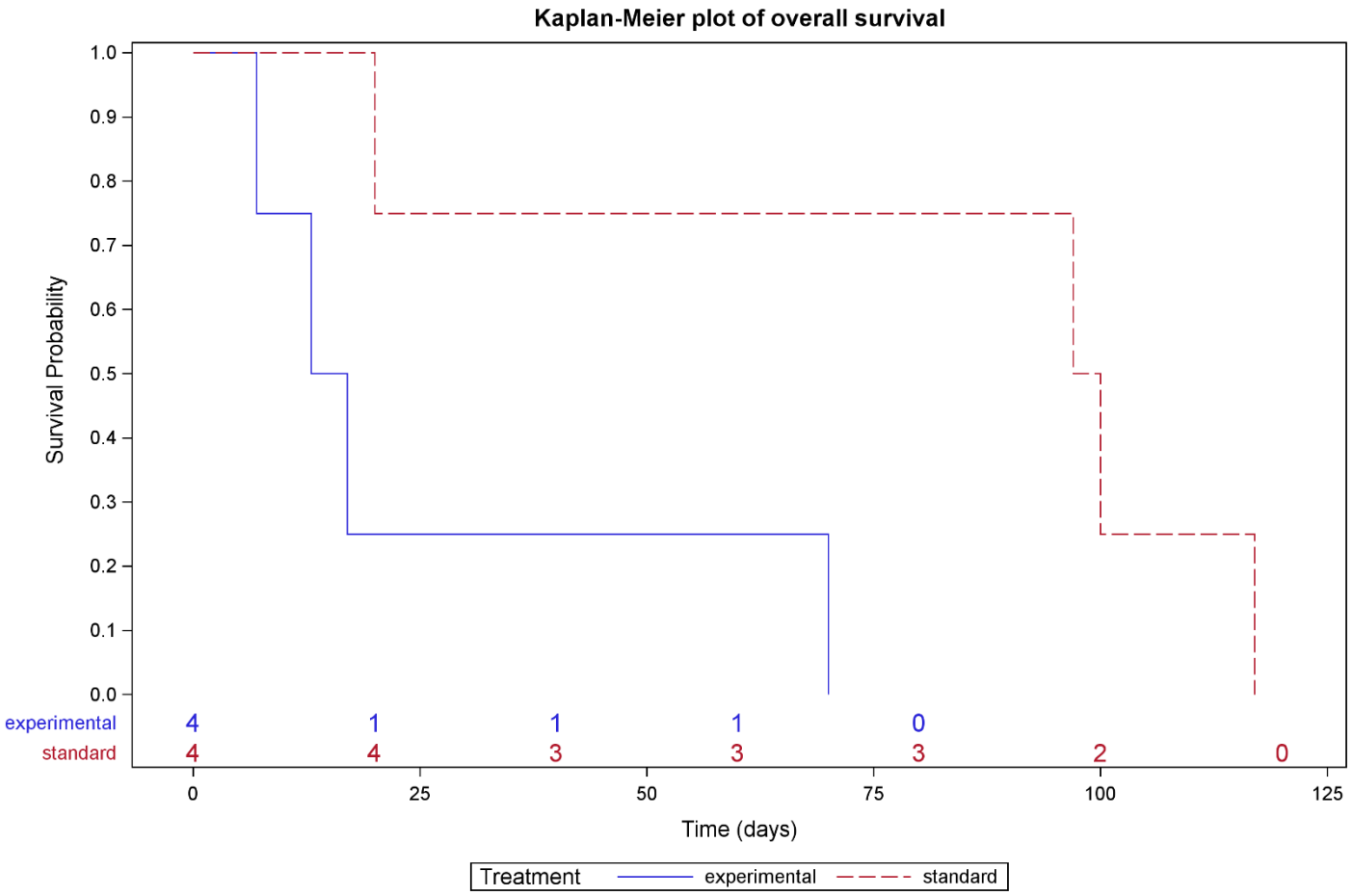
28-Jun-2018

Motivation



©HMS Analytical Software GmbH 2015

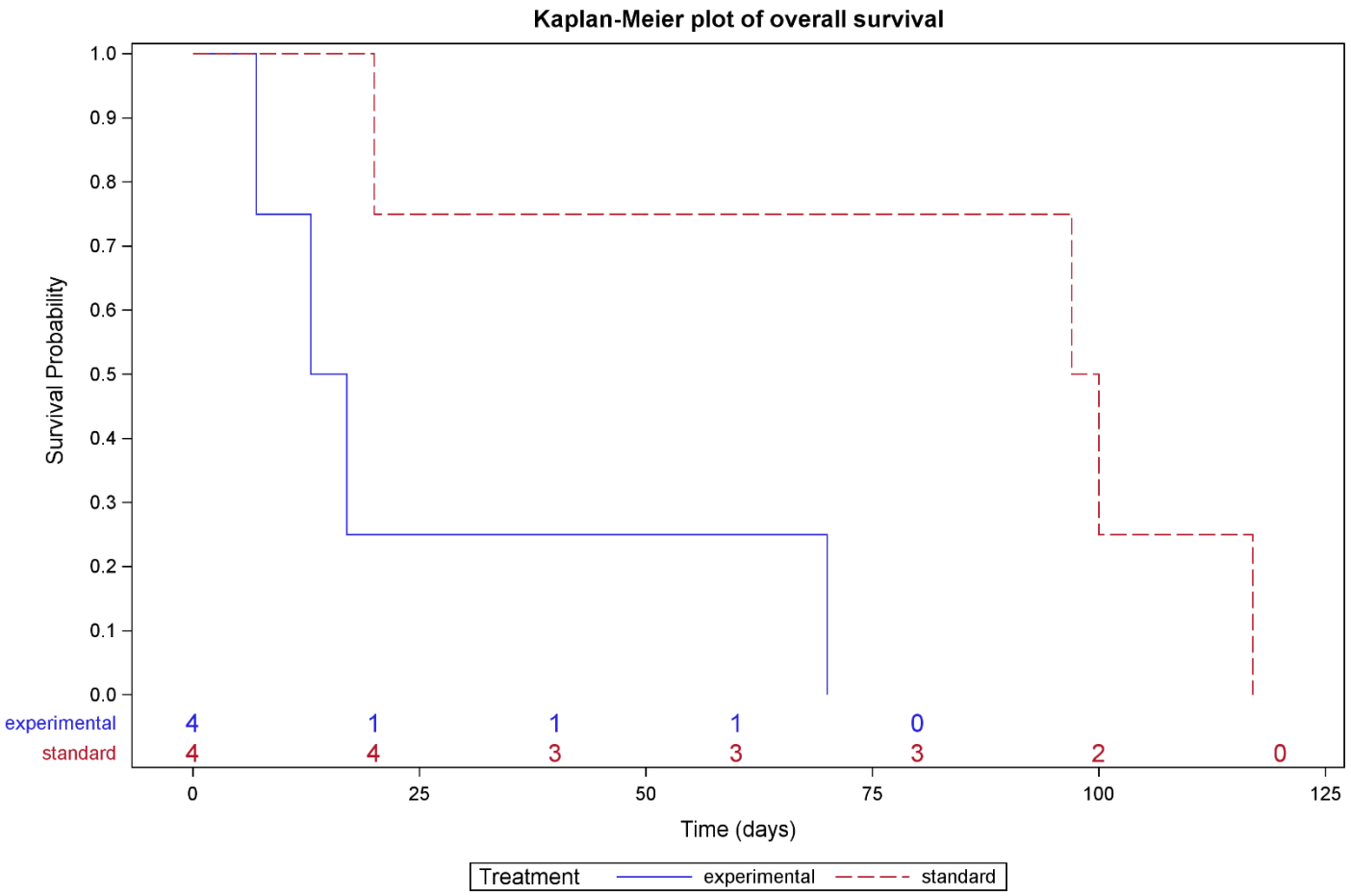
Motivation



- All patients died? 😱

©HMS Analytical Software GmbH 2015

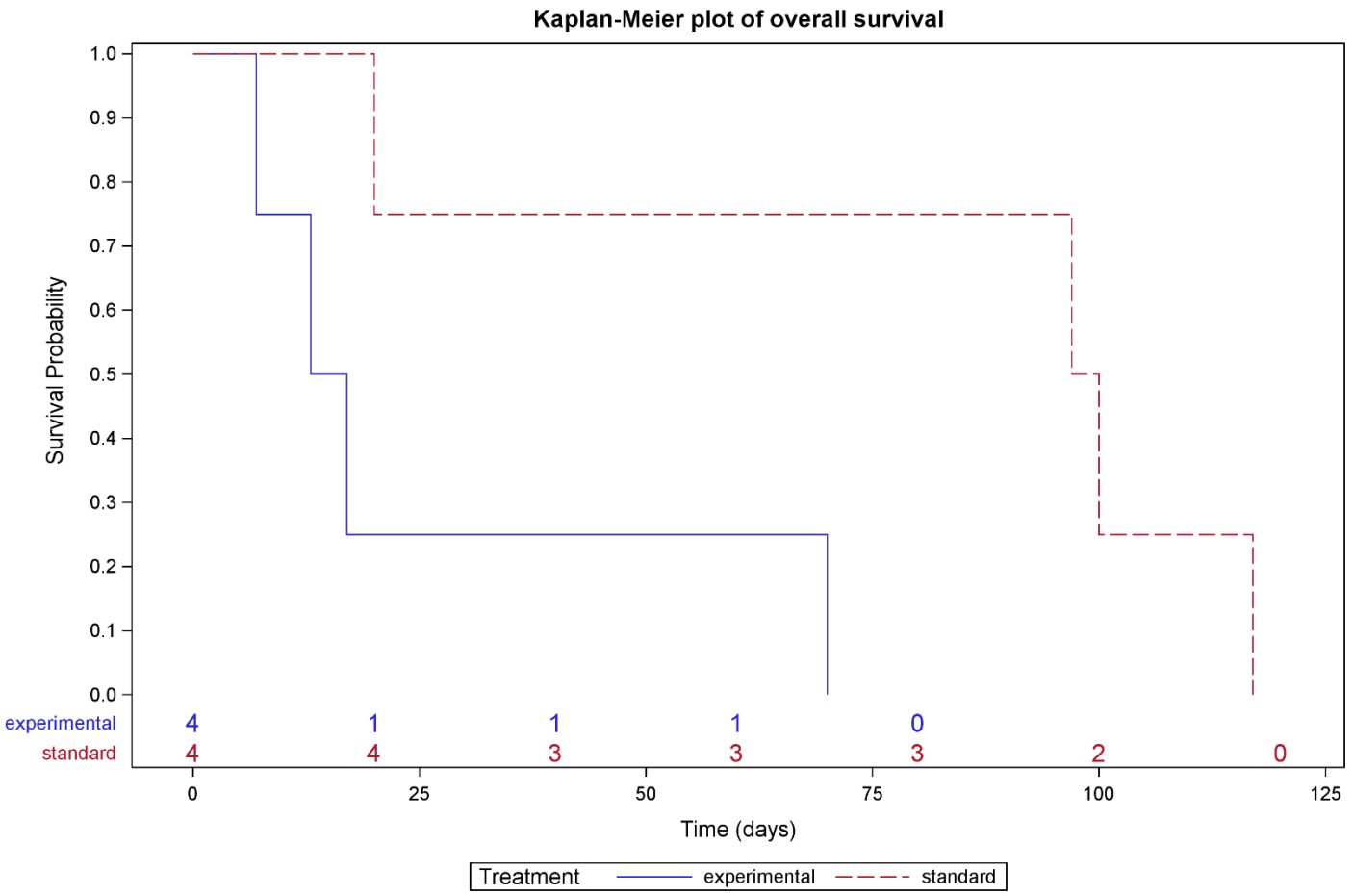
Motivation



- All patients died?
- Only 4 patients in each treatment group?

©HMS Analytical Software GmbH 2015

Motivation

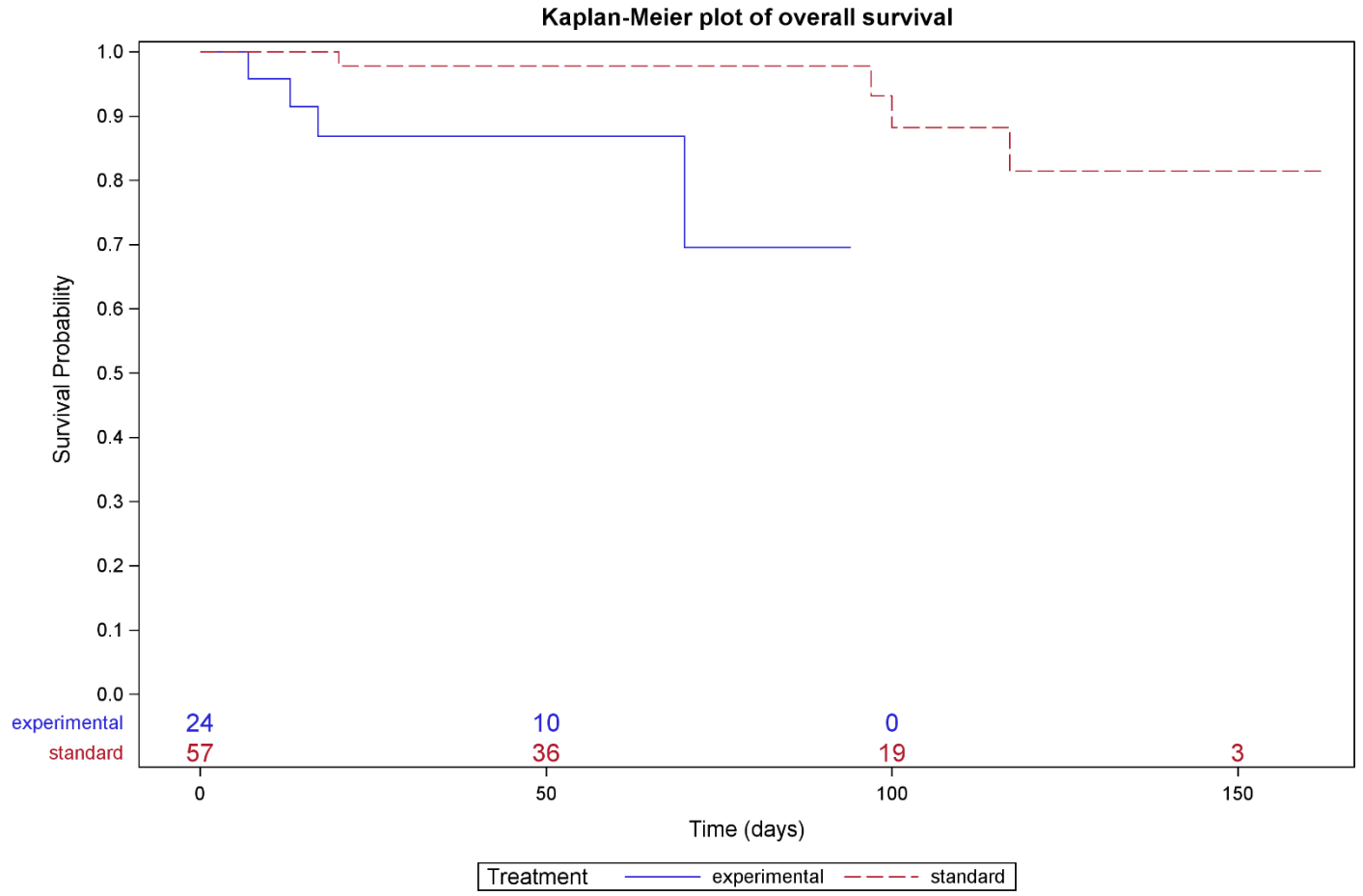


- All patients died?
- Only 4 patients in each treatment group?
- Only patients with events are included!

©HMS Analytical Software GmbH 2015

Motivation

Correct KM-plot



©HMS Analytical Software GmbH 2015

Motivation

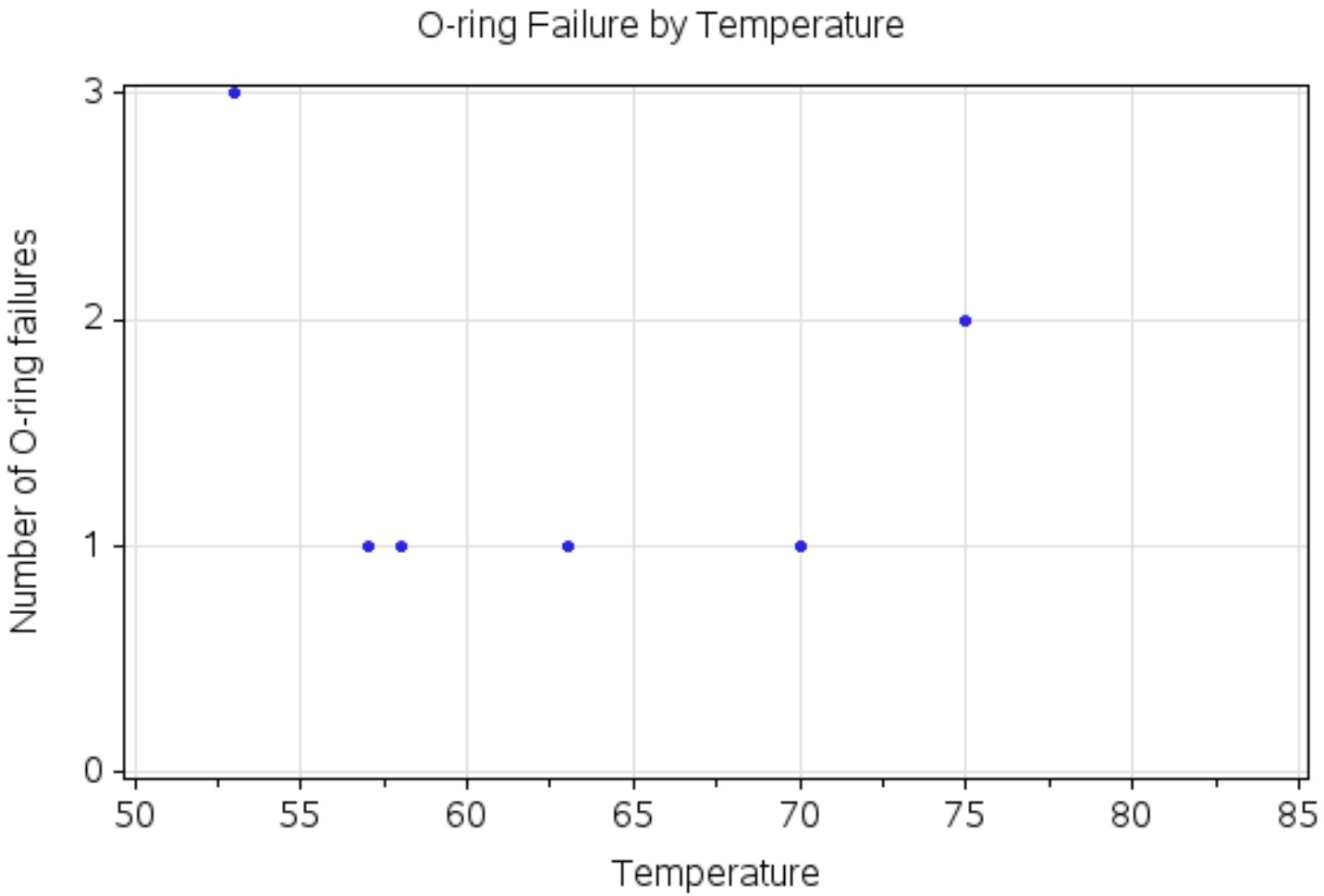
Challenger disaster

- By Kennedy Space Center - http://grin.hq.nasa.gov/ABSTRACTS/GPN-2004-00012.html, Public Domain, https://commons.wikimedia.org/w/index.php?curid=530754

©HMS Analytical Software GmbH 2015

Motivation

Challenger disaster

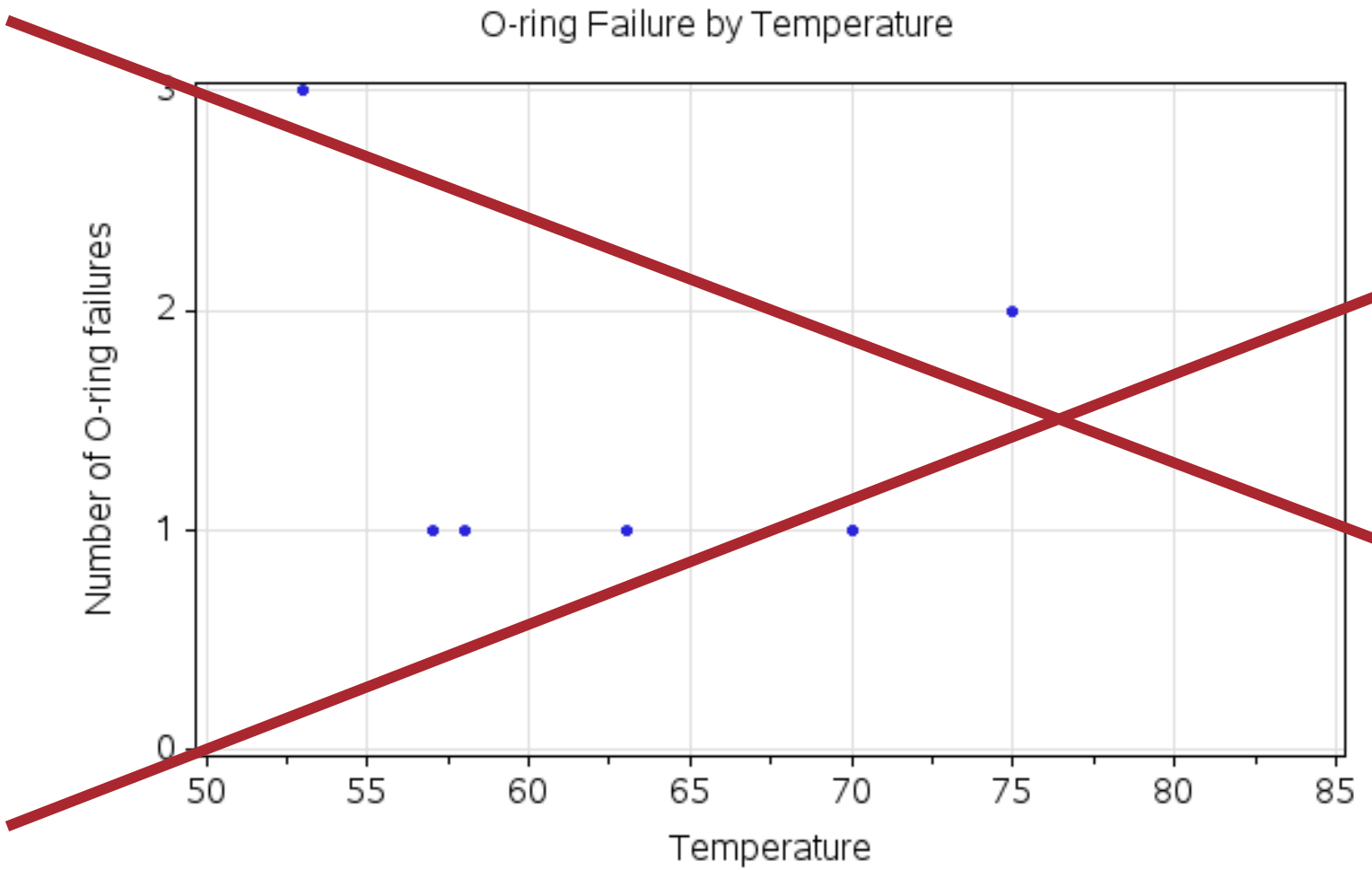


- Data analyzed before space shuttle start
- Flights with incidents of O-rings
- No relationship between temperature and failure was detected
- Thus, decision for shuttle launch

©HMS Analytical Software GmbH 2015

Motivation

Challenger disaster: wrong visualization

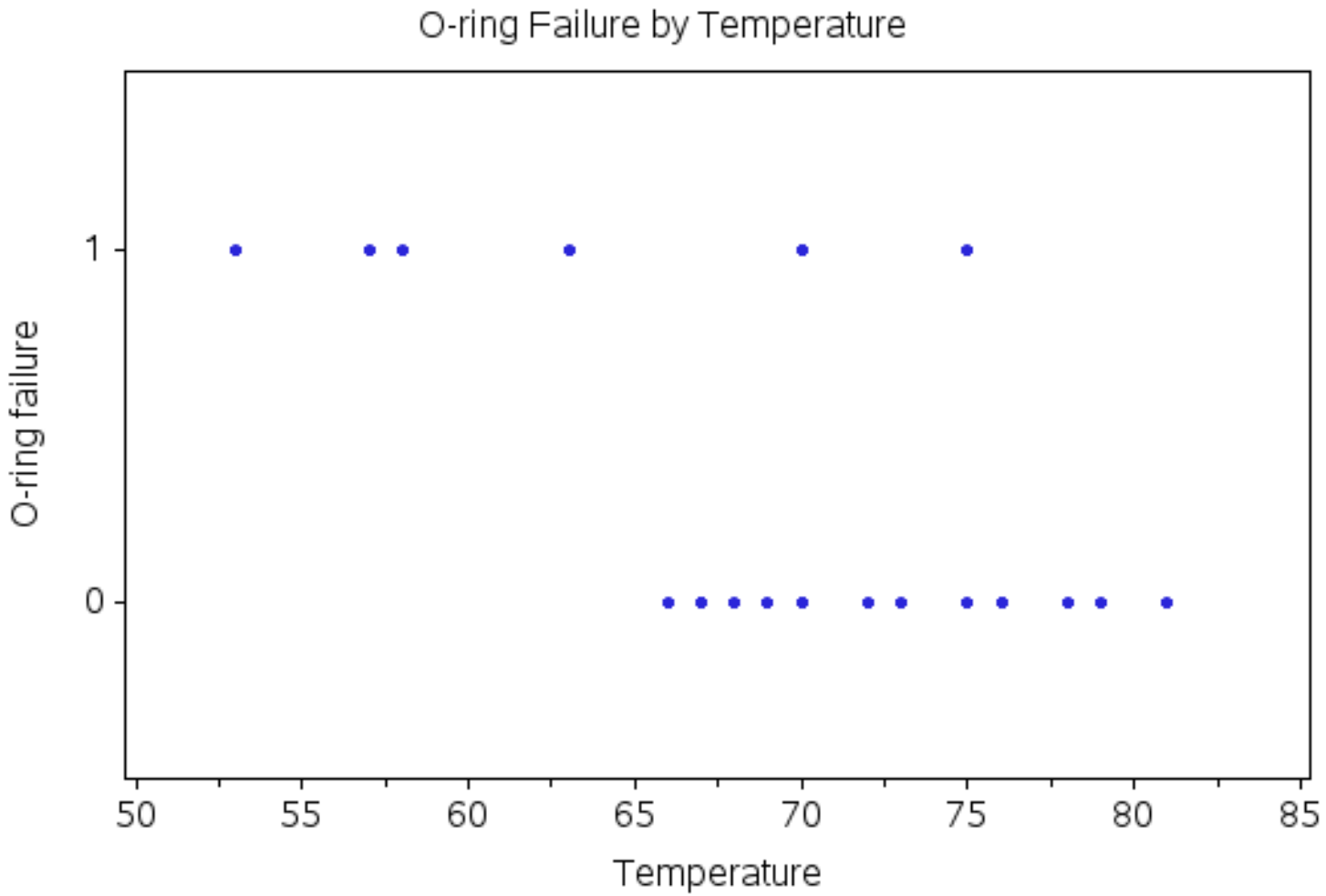


- Data analyzed before space shuttle start
- Flights with incidents of O-rings
- No relationship between temperature and failure was detected
- Thus, decision for shuttle launch

©HMS Analytical Software GmbH 2015

Motivation

Challenger disaster



- Include flights without any incidents of O-ring failure
- Relationship between temperature and failure is visible
- Temperature on the day of the launch was 31°F, so even colder than the minimum temperature in the plot
- They probably would have decided against shuttle launch

Introduction

Question: What this has to do with…

Future of Analytics?

©HMS Analytical Software GmbH 2015

Introduction

Future Analytics

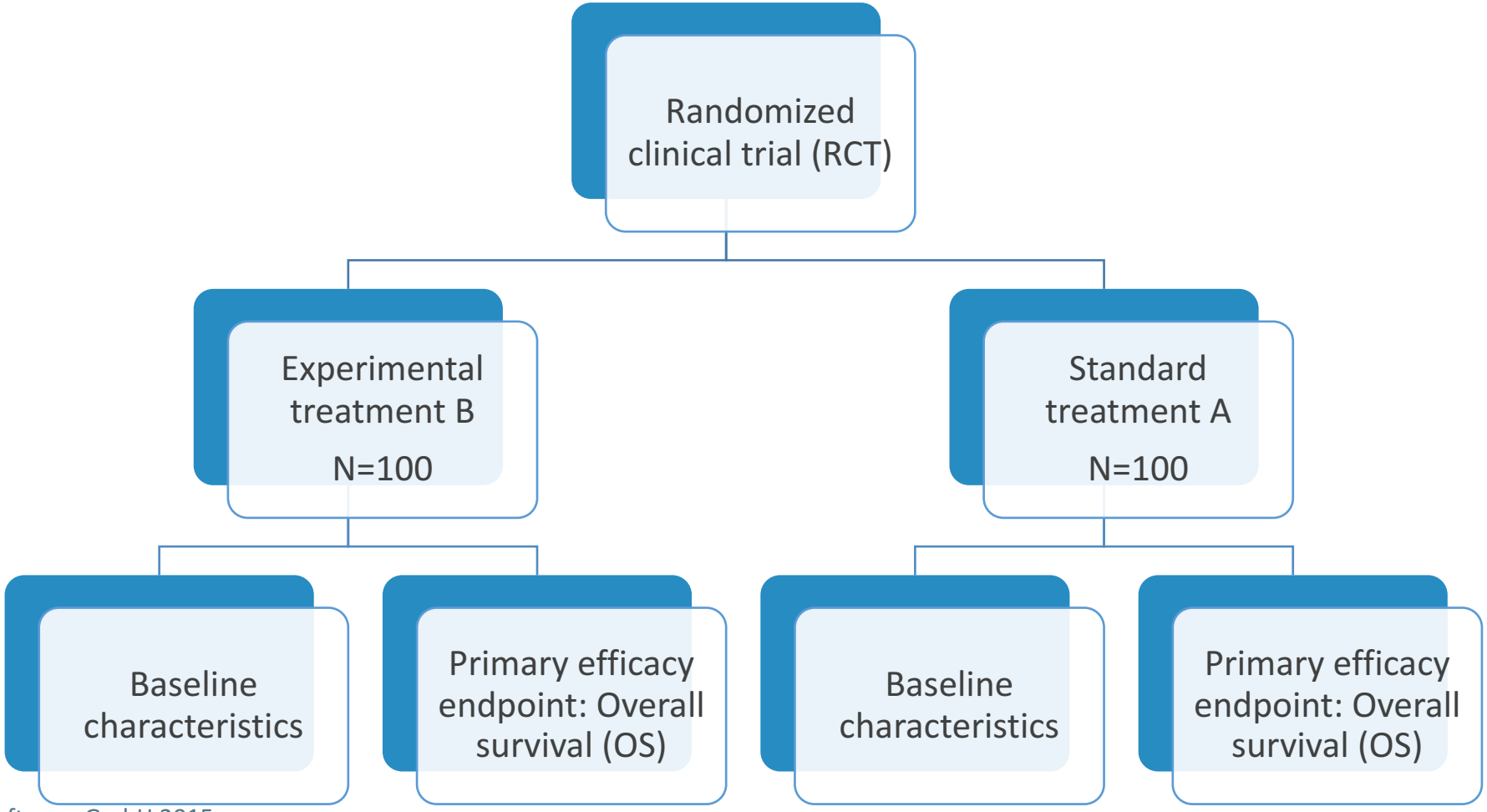
- Data visualization and new analytical tools are

©HMS Analytical Software GmbH 2015

Emoji provided free by Emoji One

Introduction

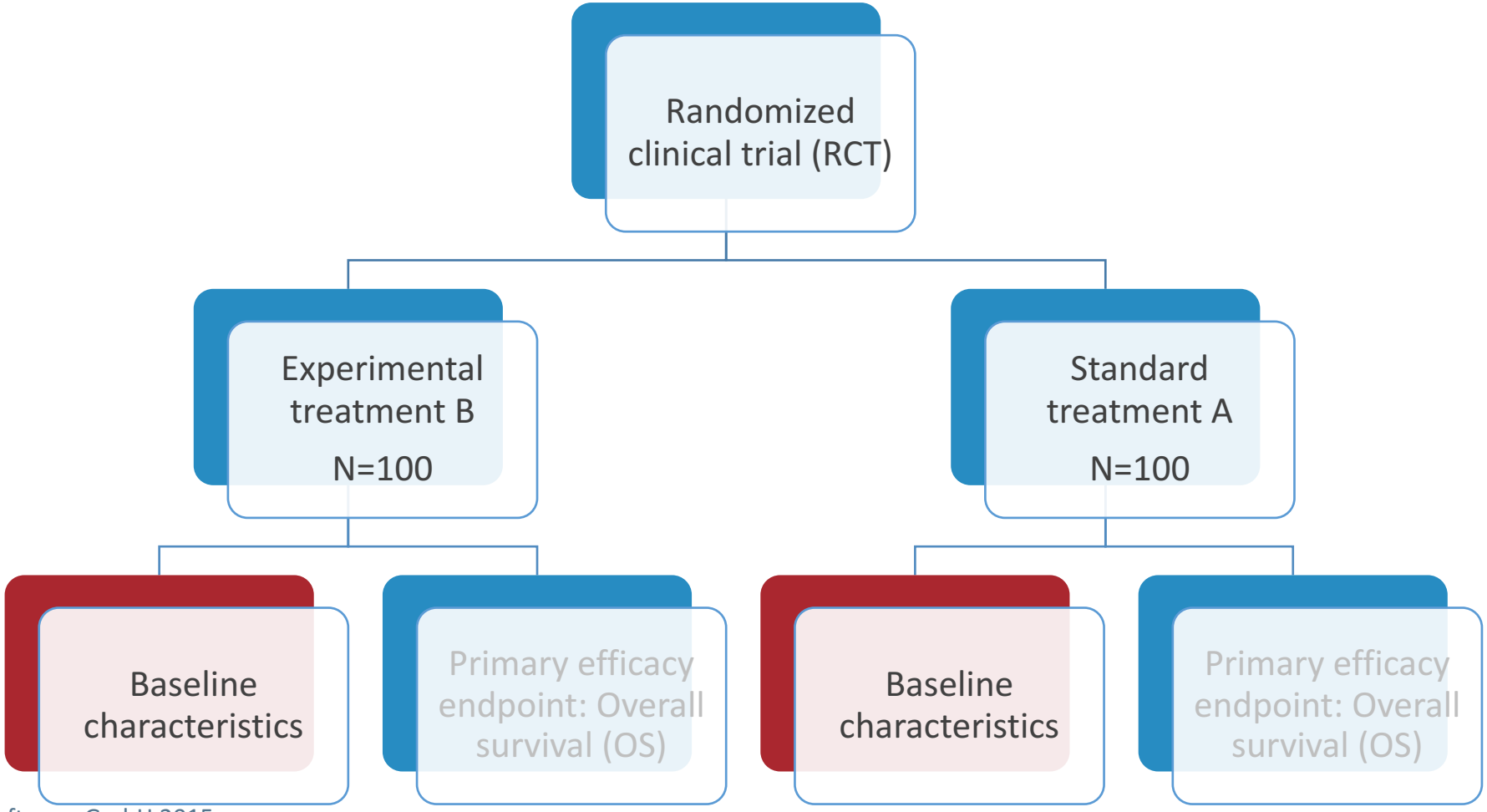
Example Study



©HMS Analytical Software GmbH 2015

First Example

Baseline characteristics



©HMS Analytical Software GmbH 2015

First Simple (Well-Known?) Example

Table of baseline characteristics

	Experimental treatment B	Standard treatment A
Number of patients, n (%)	100 (100)	100 (100)
Age		
Median (range)	62.0 (32-81)	60.5 (28-81)
Male sex, n (%)	71 (71.0)	72 (72.0)
…<other characteristics>	…	…
Bone marrow involvement, n (%)	23 (23.0)	28 (28.0)

©HMS Analytical Software GmbH 2015

First Simple (Well-Known?) Example

Table of baseline characteristics

	Experimental treatment B	Standard treatment A
Number of patients, n (%)	100 (100)	100 (100)
Age		
Median (range)	62.0 (32-81)	60.5 (28-81)
Male sex, n (%)	71 (71.0)	72 (72.0)
…<other characteristics>	…	…
Bone marrow involvement, n (%)	23 (23.0)	28 (28.0)

©HMS Analytical Software GmbH 2015

First Simple (Well-Known?) Example

Table of baseline characteristics: Wrong conclusion

	Experimental treatment B	Standard treatment A
Number of patients, n (%)	100 (100)	100 (100)
Age		
Median (range)	62.0 (32-81)	60.5 (28-81)
Male sex, n (%)	71 (71.0)	72 (72.0)
…<other characteristics>	…	…
Bone marrow involvement, n (%)	23 (23.0)	28 (28.0)

- So, 77% and 72% without bone marrow involvement?
- Biopsy results from all patients available?
- What about missing data?

©HMS Analytical Software GmbH 2015

First Simple (Well-Known?) Example

Table of baseline characteristics

	Experimental treatment B	Standard treatment A
Number of patients, n (%)	100 (100)	100 (100)
Age		
Median (range)	62.0 (32-81)	60.5 (28-81)
Male sex, n (%)	71 (71.0)	72 (72.0)
…<other characteristics>	…	…
Bone marrow involvement, n (%)	23 (23.0)	28 (28.0)

- So, 77% and 72% without bone marrow involvement?
- Biopsy results from all patients available?
- What about missing data?

©HMS Analytical Software GmbH 2015

First Simple (Well-Known?) Example

HMS Analytical Software

Table of baseline characteristics: Option 1

	Experimental treatment B	Standard treatment A
Number of patients, n (%)	100 (100)	100 (100)
Age		
Median (range)	62.0 (32-81)	60.5 (28-81)
Male sex, n (%)	71 (71.0)	72 (72.0)
…<other characteristics>	…	…
Bone marrow involvement, n (%)	23/82 (28.0)	28/83 (33.7)

- Percentages based on number of patients with non-missing data
- So, 18% and 17% of patients have missing data

©HMS Analytical Software GmbH 2015

First Simple (Well-Known?) Example

Table of baseline characteristics: Option 2

	Experimental treatment B	Standard treatment A
Number of patients, n (%)	100 (100)	100 (100)
Age		
Median (range)	62.0 (32-81)	60.5 (28-81)
Male sex, n (%)	71 (71.0)	72 (72.0)
…<other characteristics>	…	…
Bone marrow involvement*, n (%)	23 (23.0)	28 (28.0)

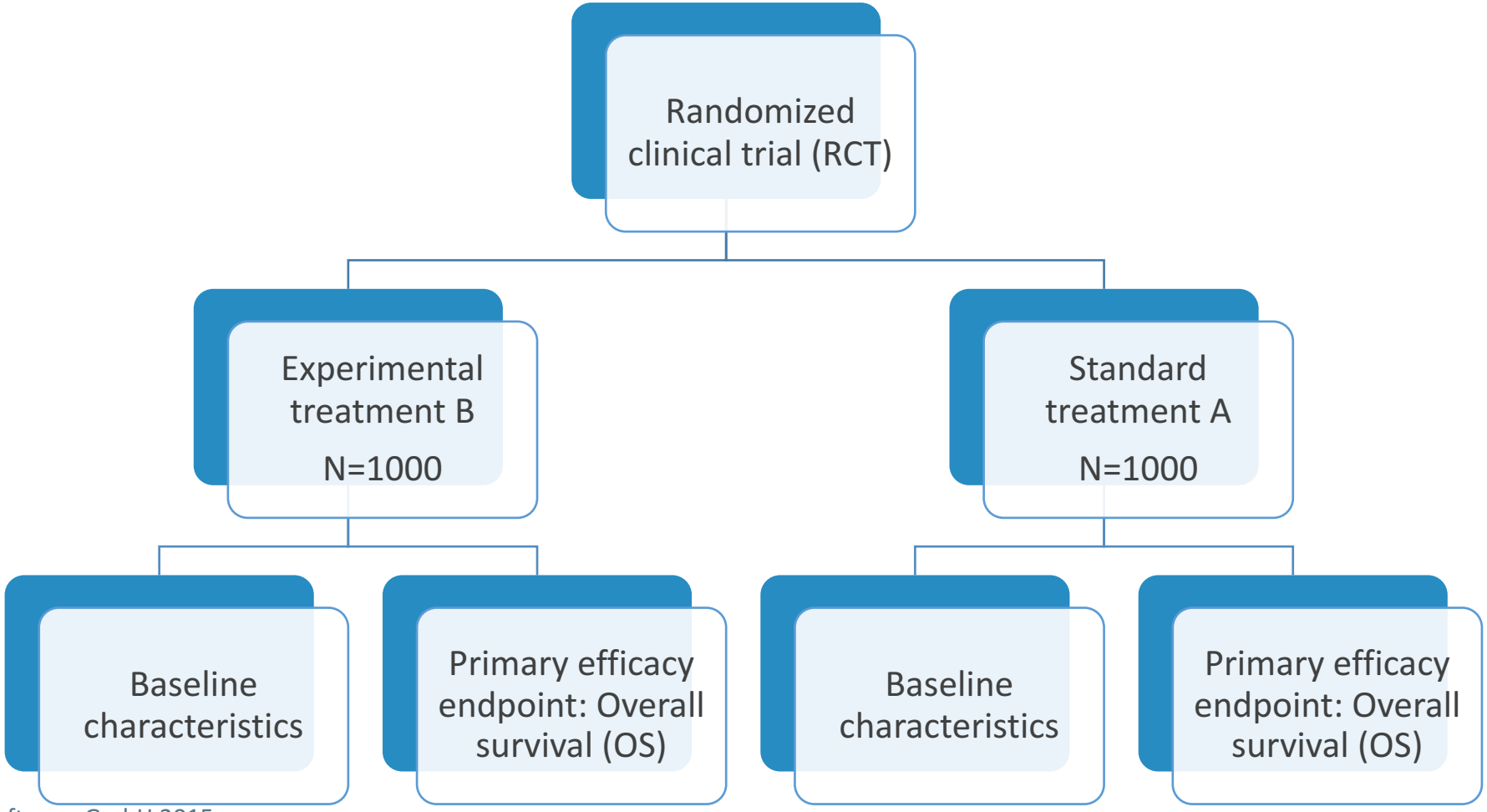
* 18 (18.0%) and 17 (17.0%) patients with missing information

- Percentages based on total number of patients
- So, 59% and 55% of patients without bone marrow involvement

©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

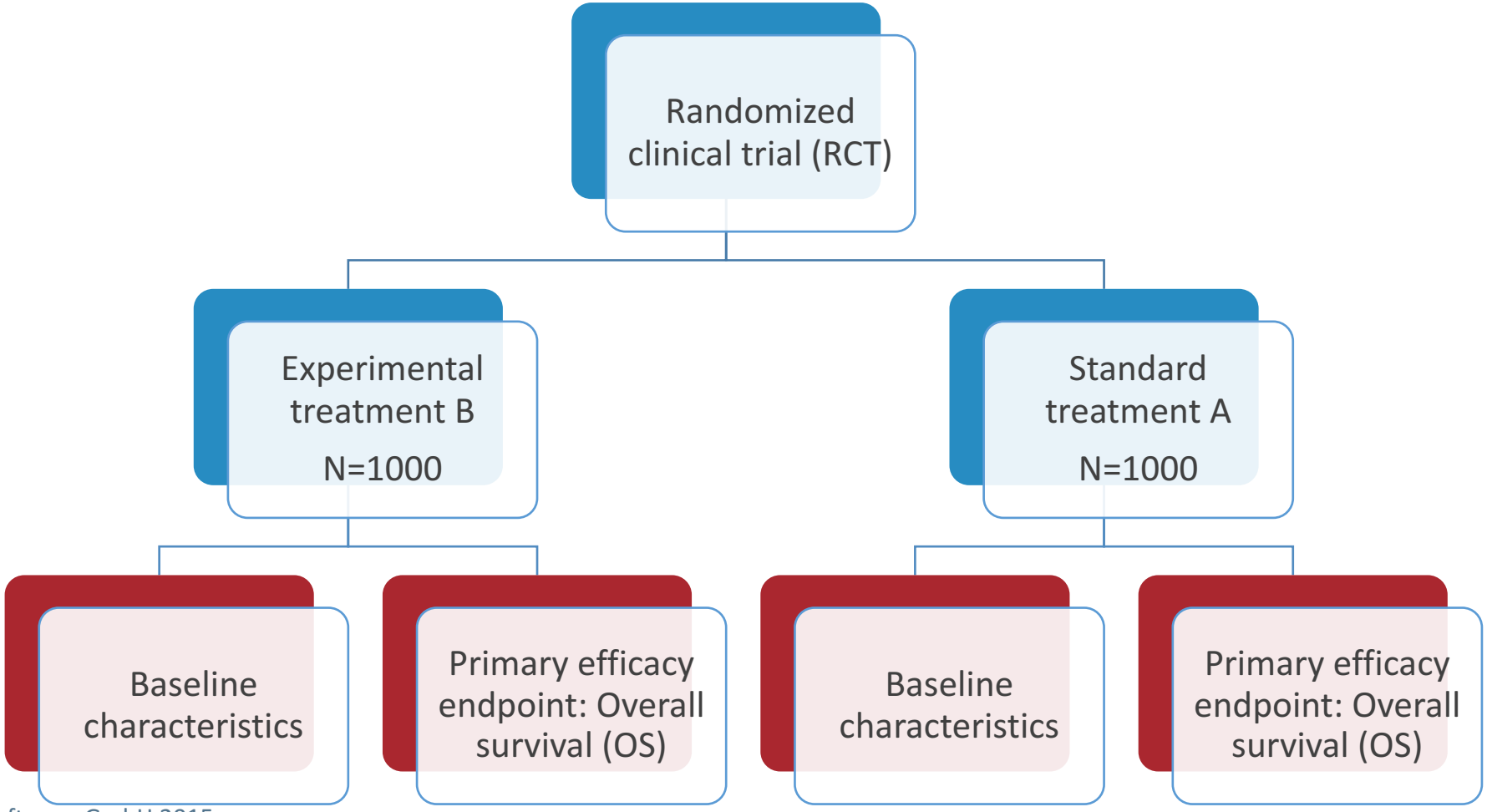
Question: Same benefit within subgroups?



©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

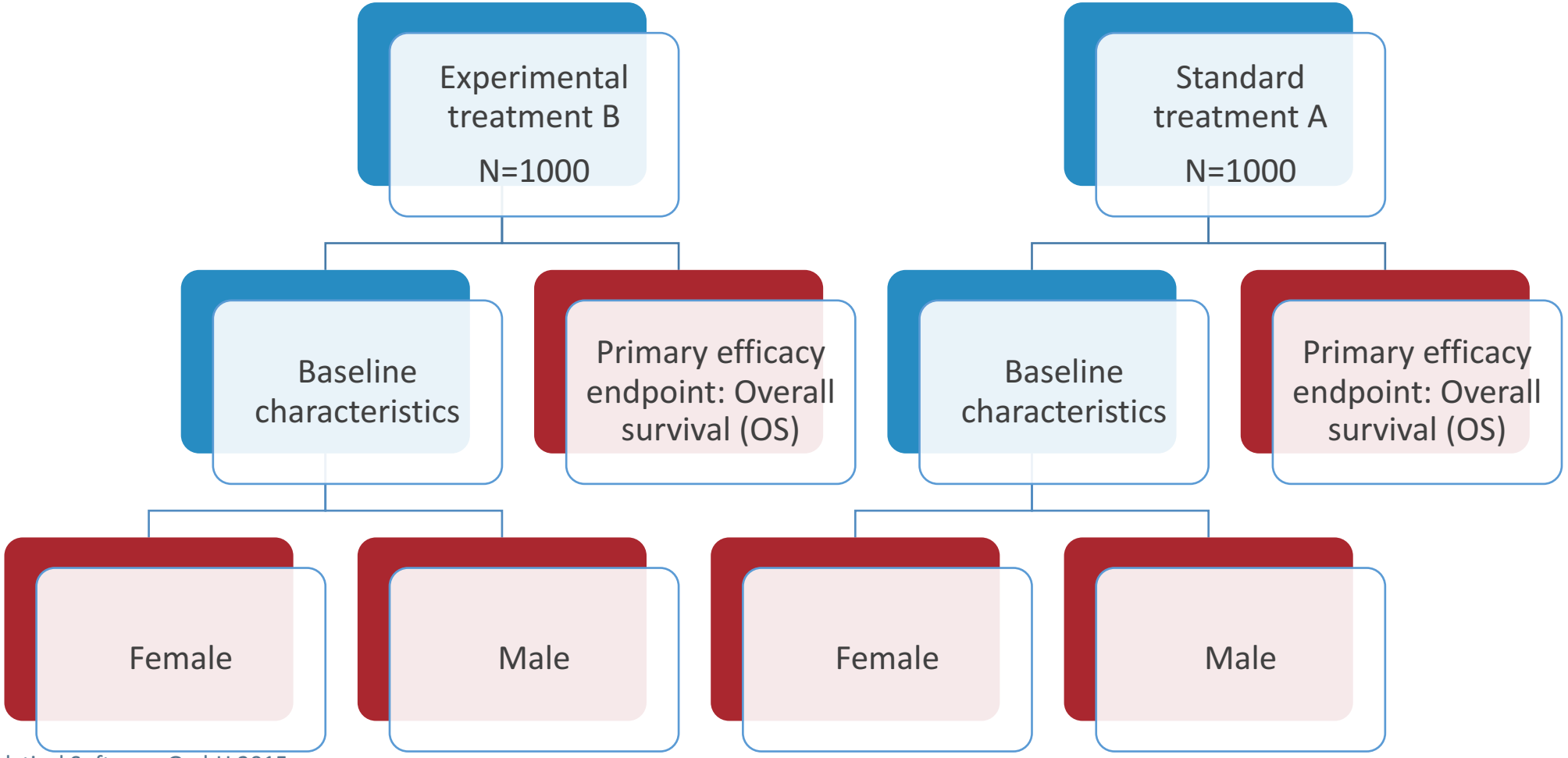
Question: Same benefit within subgroups?



©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Overall survival in female and male patients



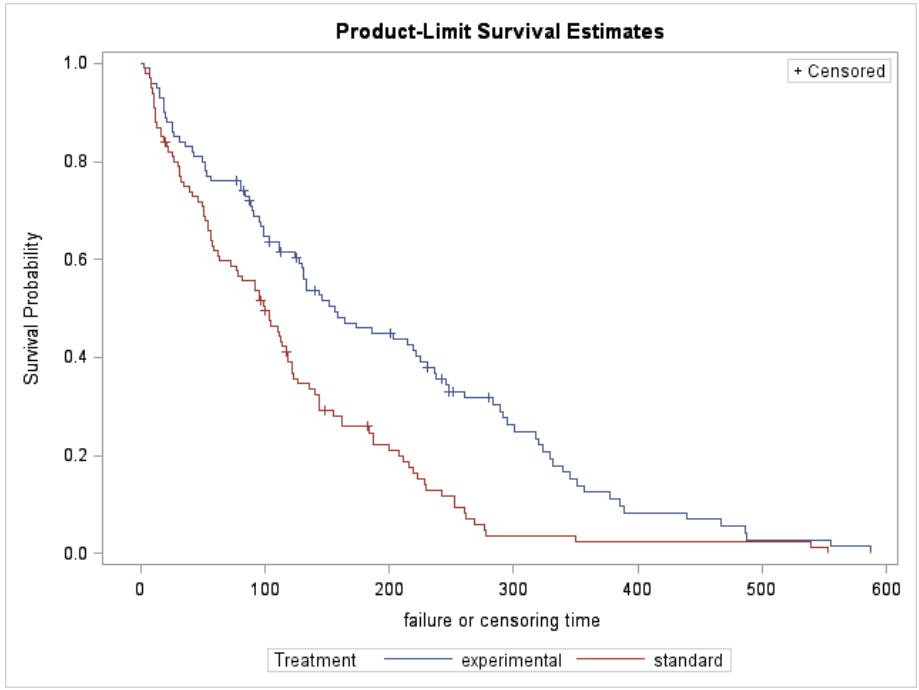
©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Main results in overall population

	Experimental treatment B	Standard treatment A
Number of patients, n (%)	100 (100)	100 (100)
Overall survival		
Number of patients with events, n (%)	87 (87.0)	94 (94.0)
Hazard ratio* (95% CI)	0.54 (0.40, 0.73)	
p-value*	<0.0001	

* from Cox proportional hazard regression model



- Benefit of experimental treatment B over standard treatment A
- Is there a consistent benefit in male and female patients?

©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Overall survival in female and male patients

	Female		Male	
	Experimental treatment B	**Standard treatment A**	**Experimental treatment B**	**Standard treatment A**
Number of patients, n (%)	29 (100)	28 (100)	71 (100)	72 (100)
Overall survival				
Number of patients with events, n (%)	25 (86.2)	26 (92.9)	62 (87.3)	68 (94.4)
Hazard ratio* (95% CI)	0.64 (0.36, 1.13)		0.51 (0.36, 0.73)	
p-value*	0.1241		0.0003	

* from Cox proportional hazard regression model

- Benefit of experimental treatment B over standard treatment A only for male patients?

©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Limitations of subgroup analyses

- Sample size much lower within the single subgroup categories and often not balanced between the categories (e.g. female prevalence lower than male prevalence)
- Study powered only for the analysis of the primary endpoint in the overall population, not for any subgroup analysis (even if pre-specified)
- Multiplicity: higher probability of incorrectly rejecting at least one null hypothesis (i.e. higher probability of incorrectly detecting a significant difference)

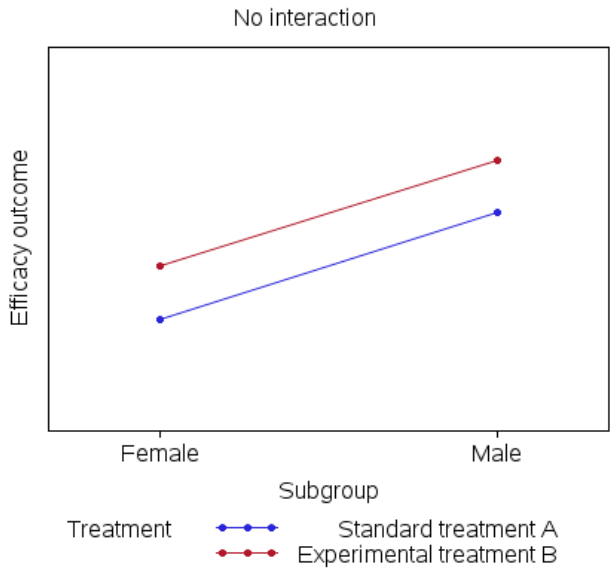
→ Better approach: Interaction test

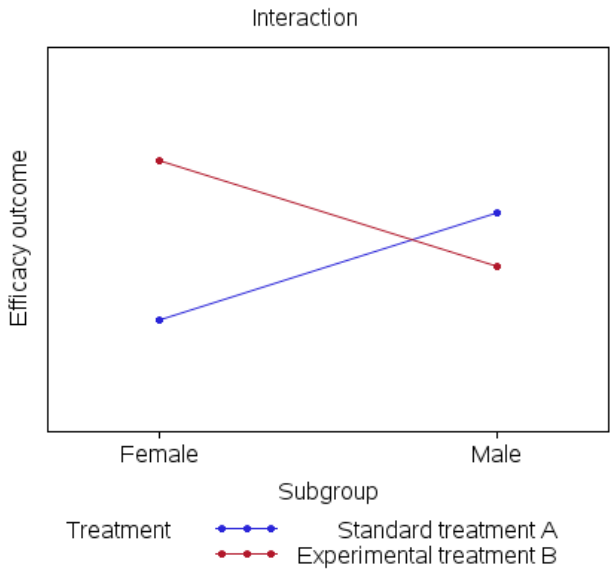
©HMS Analytical Software GmbH 2015

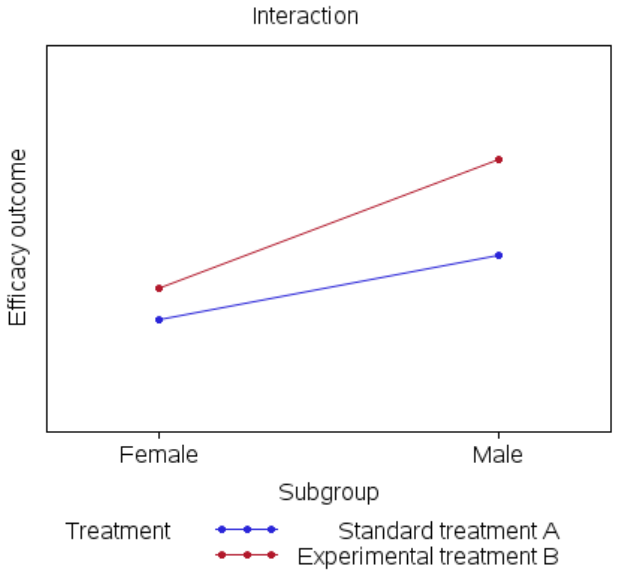
Example: Subgroup Analysis

Interaction test

- Interaction test: to check for an interaction between subgroup and treatment, i.e. if the treatment effect in one subgroup category is different to the treatment effect in the other subgroup category







©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Overall survival in female and male patients

	Female		Male	
	Experimental treatment B	**Standard treatment A**	**Experimental treatment B**	**Standard treatment A**
Number of patients, n (%)	29 (100)	28 (100)	71 (100)	72 (100)
Overall survival				
Number of patients with events, n (%)	25 (86.2)	26 (92.9)	62 (87.3)	68 (94.4)
Hazard ratio* (95% CI)	0.64 (0.36, 1.13)		0.51 (0.36, 0.73)	
Interaction p-value*	0.3612			

* from Cox proportional hazard regression model

- No significant interaction, therefore consistent benefit of experimental treatment B regardless of sex of the patient

©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Overall survival in female and male patients

	Female		Male	
	Experimental treatment B	**Standard treatment A**	**Experimental treatment B**	**Standard treatment A**
Number of patients, n (%)	29 (100)	28 (100)	71 (100)	72 (100)
Overall survival				
Number of patients with events, n (%)	25 (86.2)	26 (92.9)	62 (87.3)	68 (94.4)
Hazard ratio* (95% CI)	0.64 (0.36, 1.13)		0.51 (0.36, 0.73)	
Interaction p-value*	0.3612			

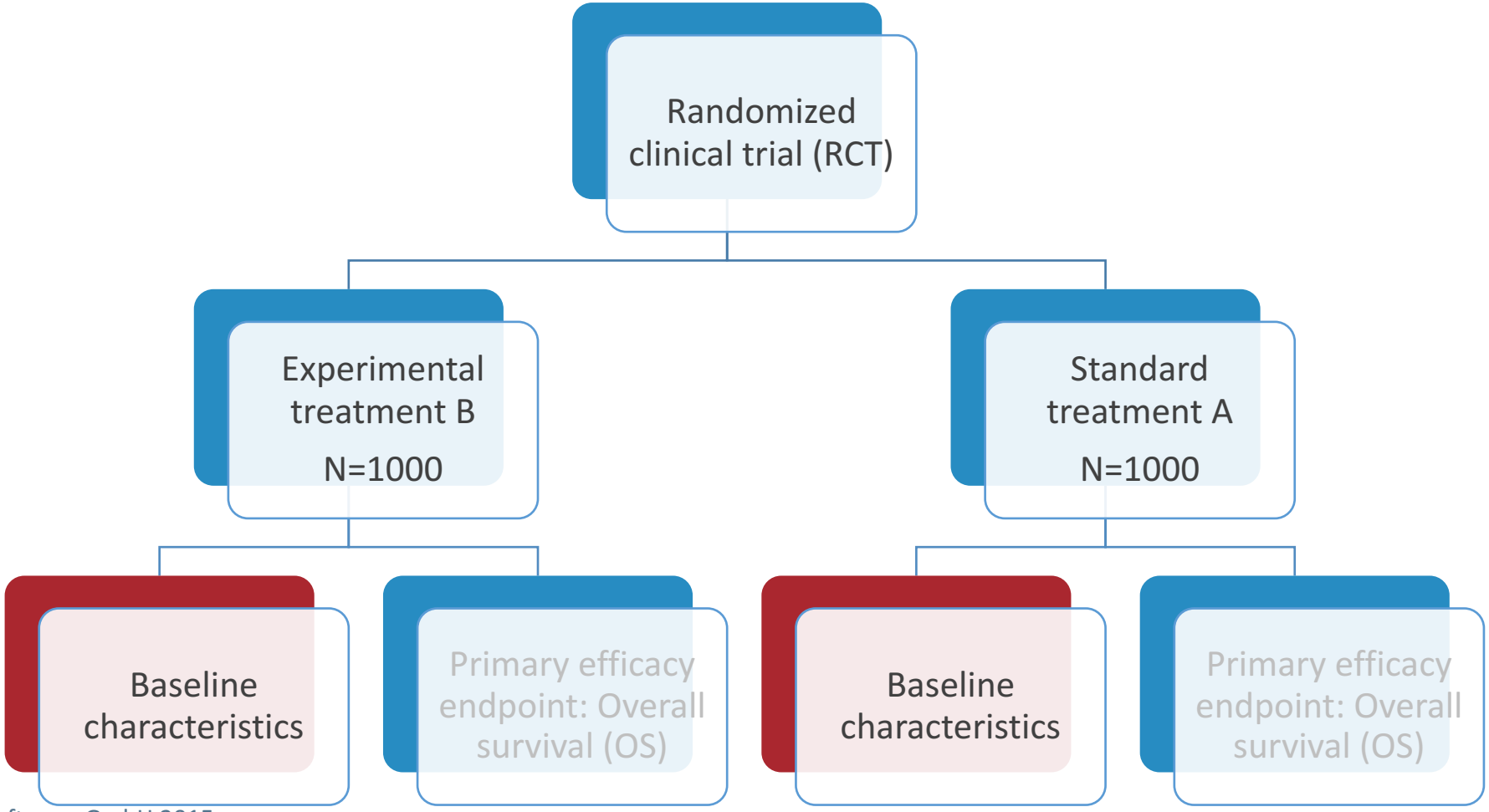
* from Cox proportional hazard regression model

- No significant interaction, therefore consistent benefit of experimental treatment B regardless of sex

©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Baseline characteristics



©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Calculating p-values for treatment comparison

- Often requested by journals
- Reviewer performed a Chi-square test, which resulted in significant p-value for a baseline variable
- Reviewer's conclusion: imbalance favored treatment A

- *Is this reasonable?*

©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Calculating p-values for treatment comparison

- If in a *randomized* trial the p-value is:

significant

- Just by chance
- Randomization went wrong

not significant

- you cannot conclude that the baseline variable did not affect the result of the primary endpoint
- Randomization has worked

©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Calculating p-values for subgroup comparison

- Requested by a journal for a subgroup analysis: are baseline characteristics balanced between subgroup categories?
- Reviewer's assumption: selection of explanatory variables for a multivariable model was based on imbalances at baseline
- Imbalances in baseline variables are possible between subgroups
  - baseline characteristics are randomized, i.e. balance between treatments but not necessarily between subgroup categories (e.g. older patients have normally more concomitant diseases)
- *Is this reasonable? → NO!*

©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Calculating p-values for subgroup comparison

- If the p-value is:

significant

- Baseline variable is not balanced between subgroup categories
- Does not say that the baseline variable is related to the outcome

not significant

- Baseline variable is balanced between subgroup categories
- Does not say that the baseline variable is not related to the outcome

©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Covariate correlation vs. covariate imbalance

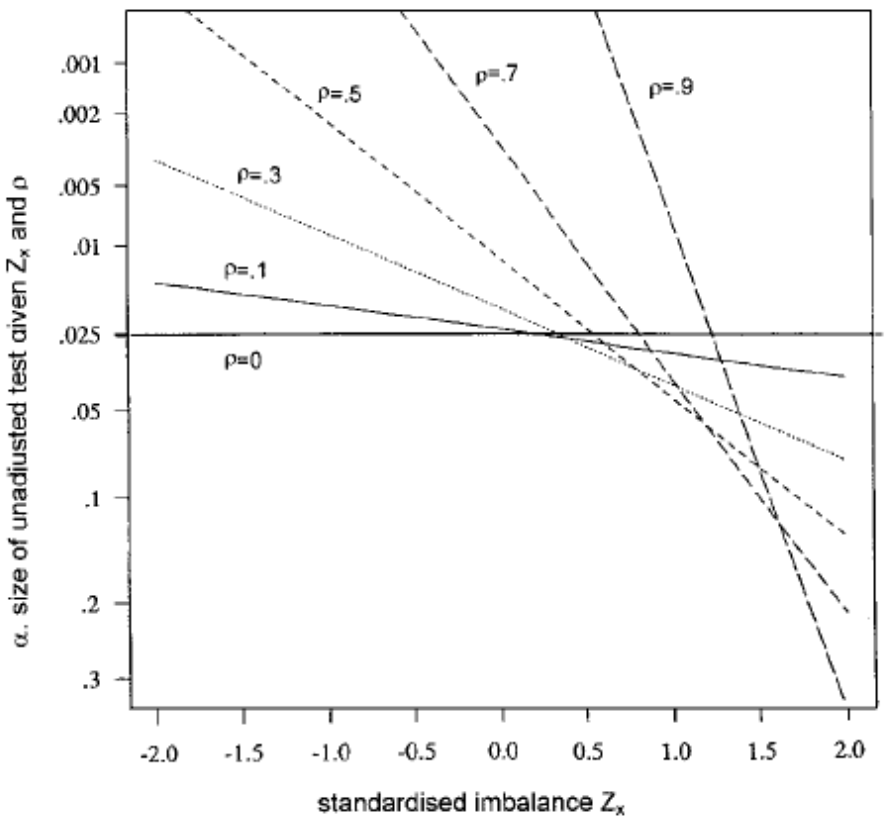


Figure 2. The effect of standardized covariate imbalance Z_x and the covariate's correlation with outcome ρ on the conditional size of unadjusted one-sided test (true $\alpha = 0.025$).

- From Pocock, S.J., Assmann, S.E., Enos, L.E. et al. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat Med, 21, pp. 2917-2930

©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Covariate correlation vs. covariate imbalance

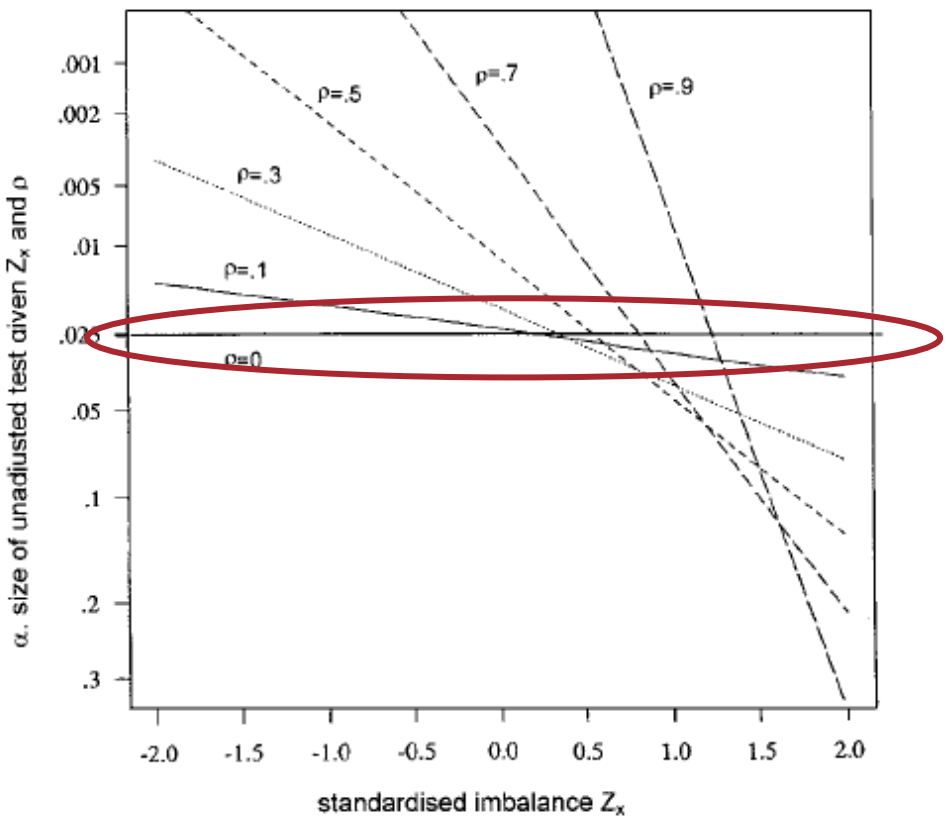


Figure 2. The effect of standardized covariate imbalance Z_x and the covariate's correlation with outcome ρ on the conditional size of unadjusted one-sided test (true $\alpha = 0.025$).

- From Pocock, S.J., Assmann, S.E., Enos, L.E. et al. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat Med, 21, pp. 2917-2930

©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Covariate correlation vs. covariate imbalance

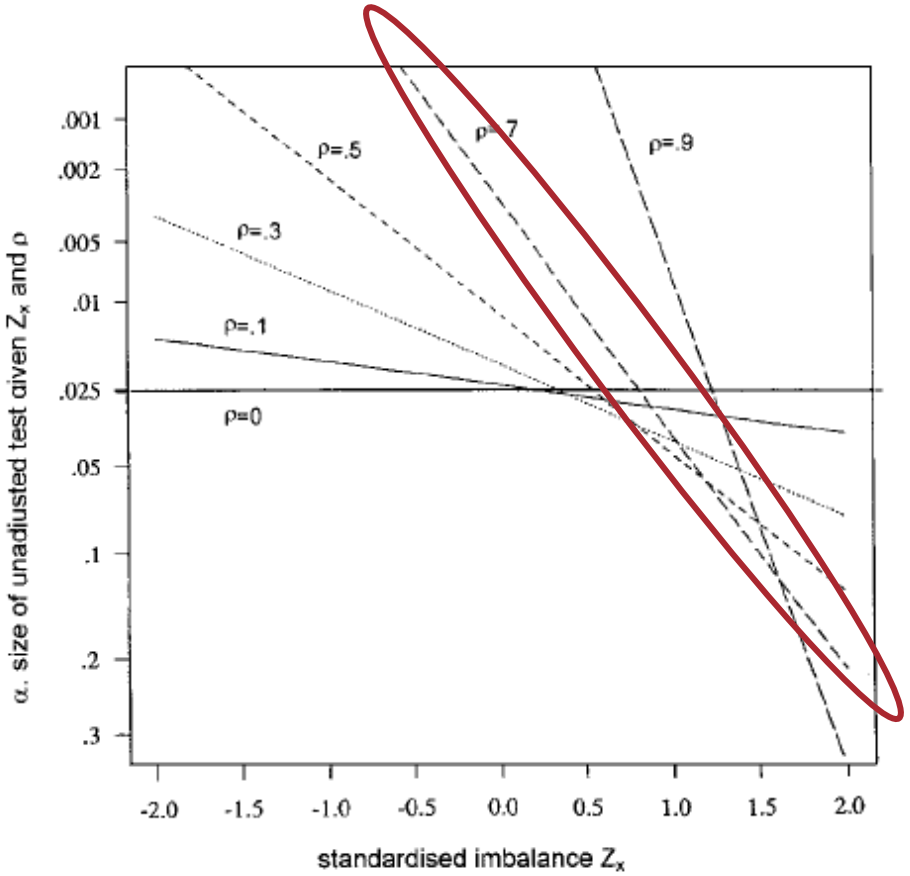


Figure 2. The effect of standardized covariate imbalance Z_x and the covariate's correlation with outcome ρ on the conditional size of unadjusted one-sided test (true $\alpha = 0.025$).

- From Pocock, S.J., Assmann, S.E., Enos, L.E. et al. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat Med, 21, pp. 2917-2930

©HMS Analytical Software GmbH 2015

Example: Comparability of Baseline Characteristics

Calculating p-values for treatment comparison

Does not make sense in a randomized trial

Tests whether the treatment groups have been randomly assigned

In case of balance between the groups no conclusion possible that baseline variable has no impact on the result of the (primary) endpoint

In case of imbalance between the groups no conclusion possible whether the baseline variable may have an impact on the result of the (primary) endpoint

©HMS Analytical Software GmbH 2015

Summary

HMS Analytical Software

Please…

- Don’t believe everything you see
- Check for handling of missing values
- Check if data is appropriately visualized
- Check if all data is included
- Check sample size (relevance of differences)
- Do an interaction test when performing subgroup analyses
- Use words like “reasonable”, “considerable”, “relevant” instead of “significant” in case of exploratory p-values
- Use p-values sparingly – look at confidence intervals instead

©HMS Analytical Software GmbH 2015

Conclusion

Data visualization and new analytical tools

Does this mean:

- Efficient = less CPU usage
- Fast = takes few seconds to create e.g. a plot
- Interactive = tool immediately responds to the user's input
- Smart = knows what to do
- Friendly = nice user-interface
- Secure = validated tool
- Powerful = can do "everything"
- Fresh = delivers results without delay
- Easy = everybody can use it

©HMS Analytical Software GmbH 2015

Conclusion

Data visualization and new analytical tools

Does this mean:

- Efficient = less CPU usage
- Fast = takes few seconds to create e.g. a plot
- Interactive = tool immediately responds to the user's input
- Smart = knows what to do
- Friendly = nice user-interface
- Secure = validated tool
- Powerful = can do "everything"
- Fresh = delivers results without delay
- Easy = everybody can use it

©HMS Analytical Software GmbH 2015

Conclusion

HMS Analytical Software

Data visualization and new analytical tools

Does this mean:

- Efficient = less CPU usage
- **Fast = takes few seconds to create e.g. a plot**
- **Interactive = tool immediately responds to the user's input**
- Smart = knows what to do
- Friendly = nice user-interface
- Secure = validated tool
- Powerful = can do "everything"
- **Fresh = delivers results without delay**
- Easy = everybody can use it

©HMS Analytical Software GmbH 2015

Conclusion

Data visualization and new analytical tools

Does this mean:

- Efficient = less CPU usage
- Fast = takes few seconds to create e.g. a plot
- Interactive = tool immediately responds to the user's input
- **Smart = knows what to do**
- Friendly = nice user-interface
- **Secure = validated tool**
- **Powerful = can do "everything"**
- Fresh = delivers results without delay
- **Easy = everybody can use it**

©HMS Analytical Software GmbH 2015

Conclusion

Data visualization and new analytical tools

What is still necessary:

- Validation of results
- Validation of underlying dataset
- (Statistical) output review
- Analysis plan: select appropriate statistical methods

→ Raise people's awareness of these essential facts

©HMS Analytical Software GmbH 2015

References

- Altman, D. G. (1985) Comparability of randomized groups, *The Statistician,* 34, pp. 125-136
- Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989) Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure, *Journal of the American Statistical Association,* 84(408), pp. 945-957
- Dmitrienko, A., Muysers, C., Fritsch, A. and Lipkovich, I. (2016) General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials, *Journal of Biopharmaceutical Statistics*, 26:1, pp. 71-98
- European Medicines Agency: EMA/CHMP/295050/2013 – Guideline on adjustment for baseline covariates in clinical trials. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500184923.pdf, February 2015
- European Medicines Agency: EMA/CHMP/539146/2013 – Guideline on the investigation of subgroups in confirmatory clinical trials, DRAFT http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf, January 2014
- Pocock, S.J., Assmann, S.E., Enos, L.E. et al. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, *Stat Med*, 21, pp. 2917-2930
- Senn, S. (1994) Testing for baseline balance in clinical trials, *Statistics in Medicine*, pp. 1715-1726

©HMS Analytical Software GmbH 2015

THANK YOU!

Any questions?

©HMS Analytical Software GmbH 2015

Contact information

Corinna Miede

Principal Statistician

+49- 6221 6051 196
Corinna.Miede@analytical-software.de
Huteweg 4, 35096 Weimar (Lahn), Germany

HMS Analytical Software GmbH
Rohrbacher Str. 26
69115 Heidelberg
www.analytical-software.de
info@analytical-software.de
+49-6221-6051-0

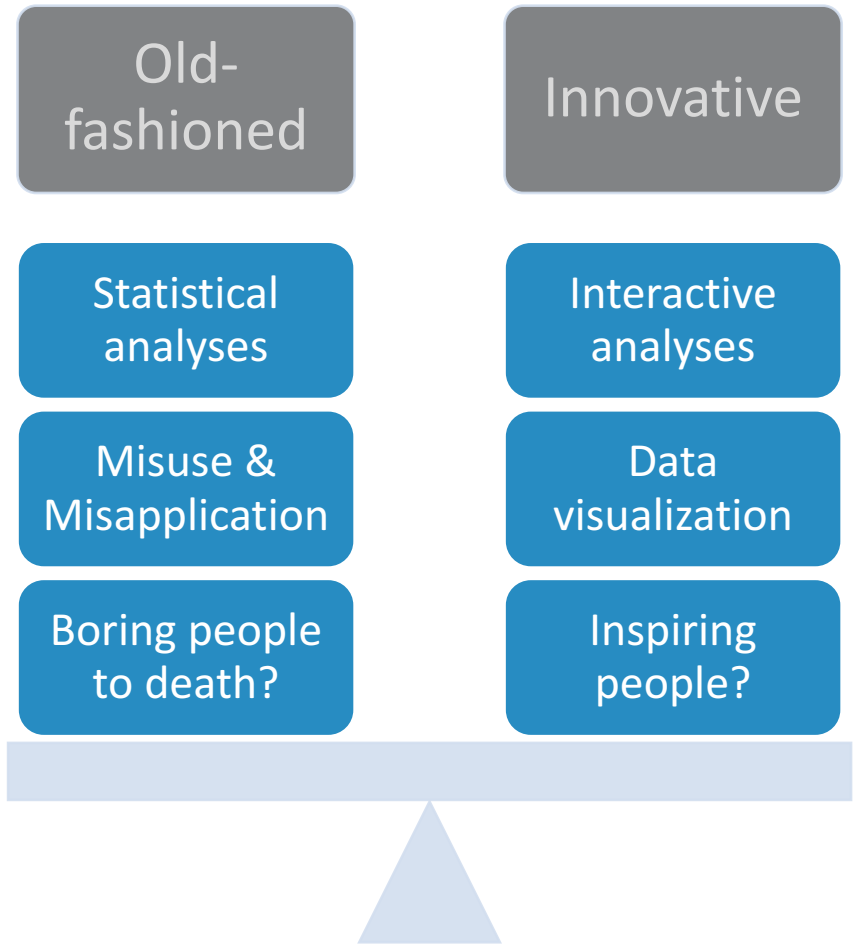
XING HMS auf XING: https://www.xing.com/company/hmsanalyticalsoftwaregmbh

Back-up Slides

©HMS Analytical Software GmbH 2015

Introduction

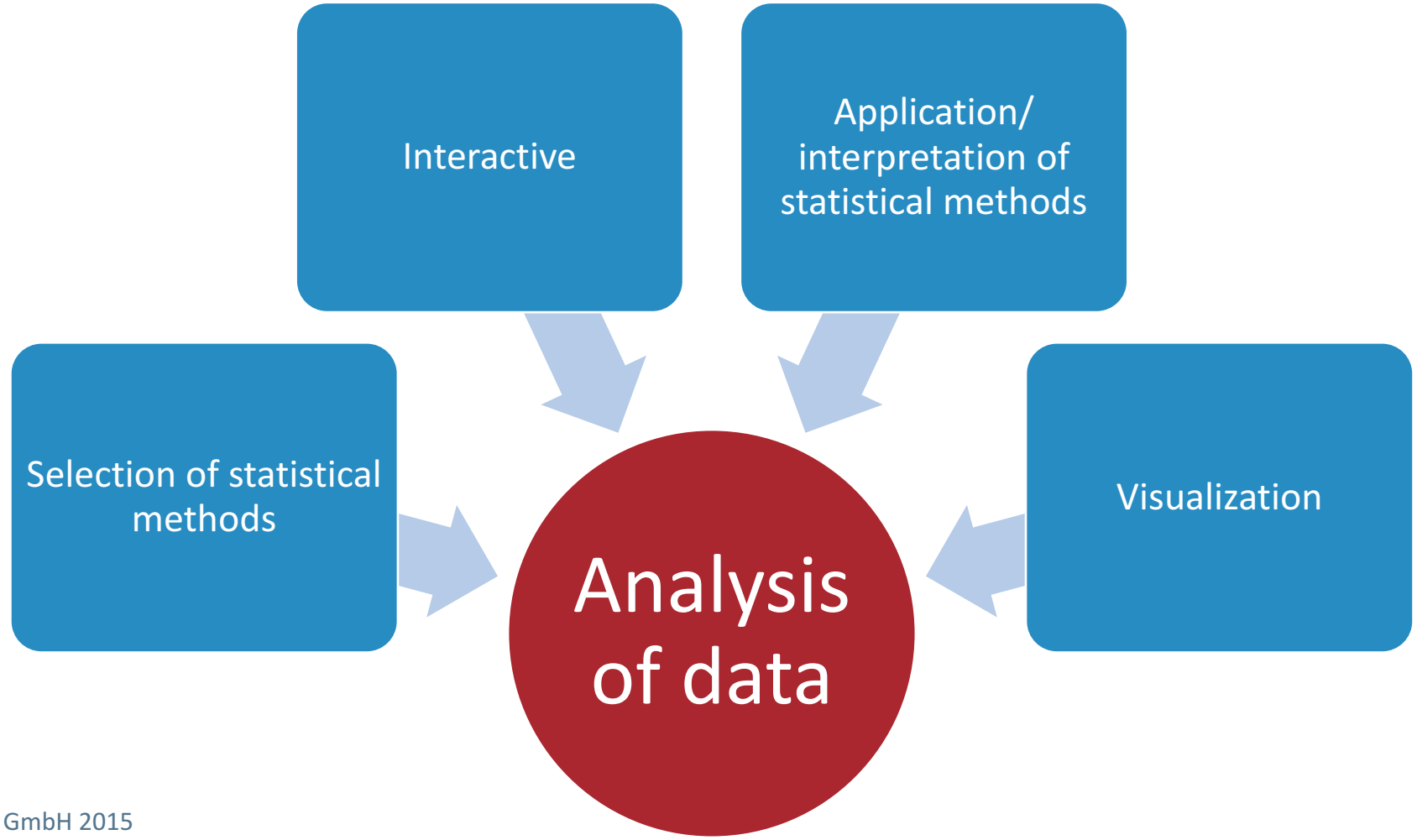
Question: Old-fashioned vs. Innovative?



©HMS Analytical Software GmbH 2015

Introduction

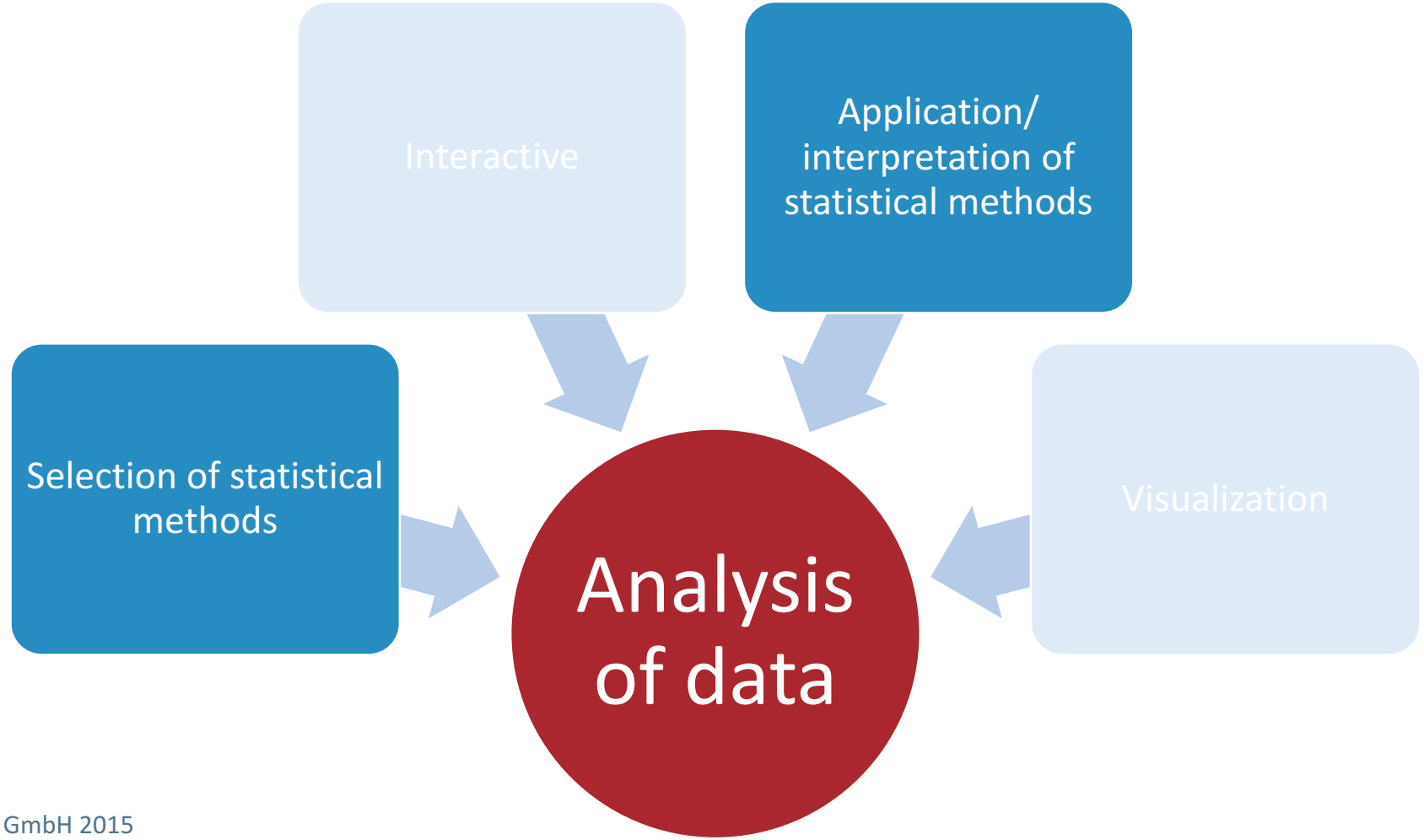
Answer: NO!



©HMS Analytical Software GmbH 2015

Introduction

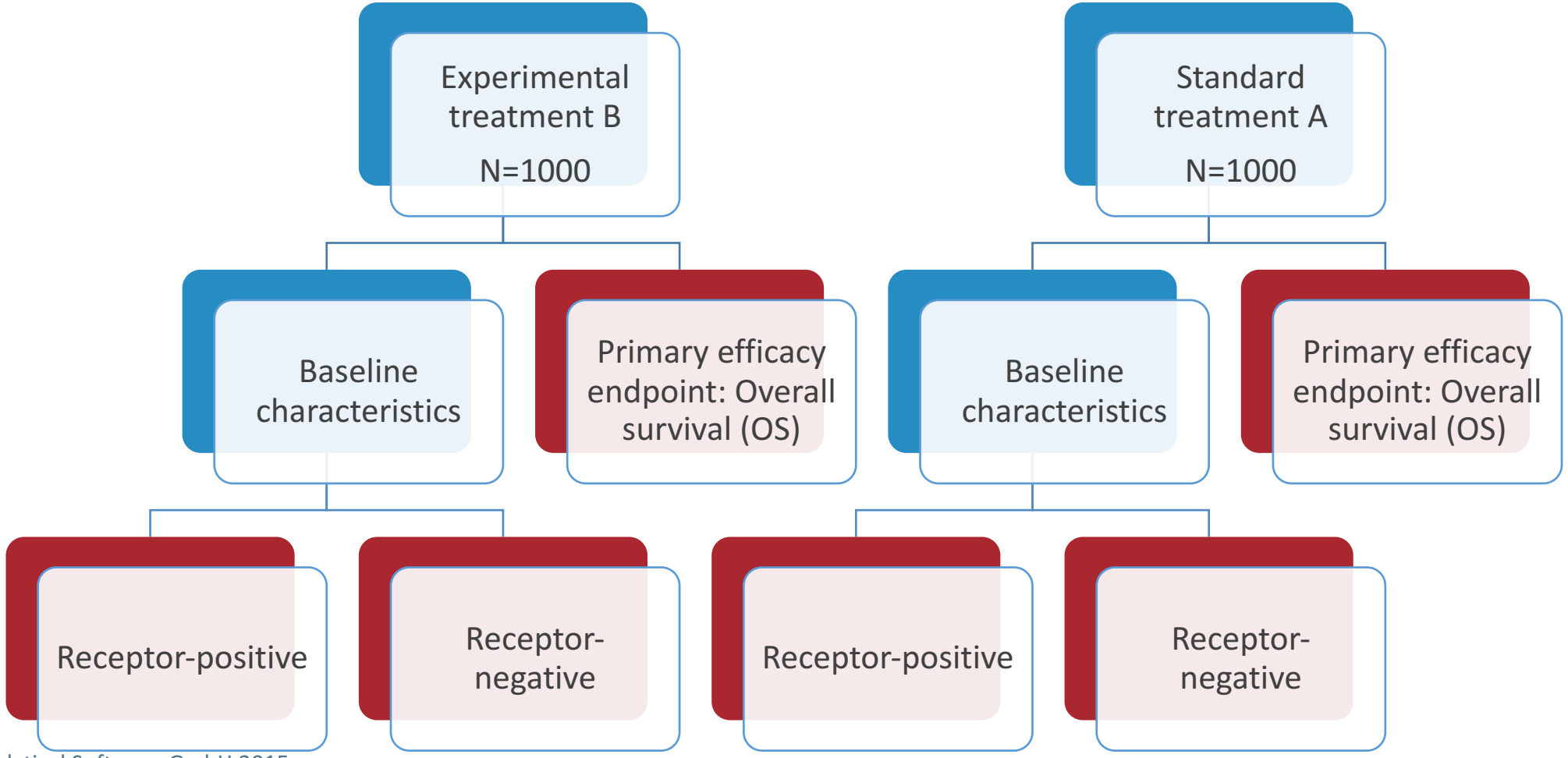
Focus of this presentation



©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Overall survival in receptor-positive and receptor-negative patients



©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Main results in overall population

	Experimental treatment B	Standard treatment A
Number of patients, n (%)	100 (100)	100 (100)
Overall survival		
Number of patients with events, n (%)	87 (87.0)	94 (94.0)
Hazard ratio* (95% CI)	0.85 (0.63, 1.14)	
p-value*	0.2692	

* from Cox proportional hazard regression model

- No significant benefit of experimental treatment B over standard treatment A
- Any benefit in specific subgroups?

©HMS Analytical Software GmbH 2015

Example: Subgroup Analysis

Overall survival in receptor-positive and receptor-negative patients

	Receptor-negative		Receptor-positive	
	Experimental treatment B	**Standard treatment A**	**Experimental treatment B**	**Standard treatment A**
Number of patients, n (%)	68 (100)	66 (100)	32 (100)	34 (100)
Overall survival				
Number of patients with events, n (%)	63 (92.6)	61 (92.4)	24 (75.0)	33 (97.1)
Hazard ratio* (95% CI)	1.11 (0.78, 1.59)		0.41 (0.23, 0.72)	
Interaction p-value*	0.0057			

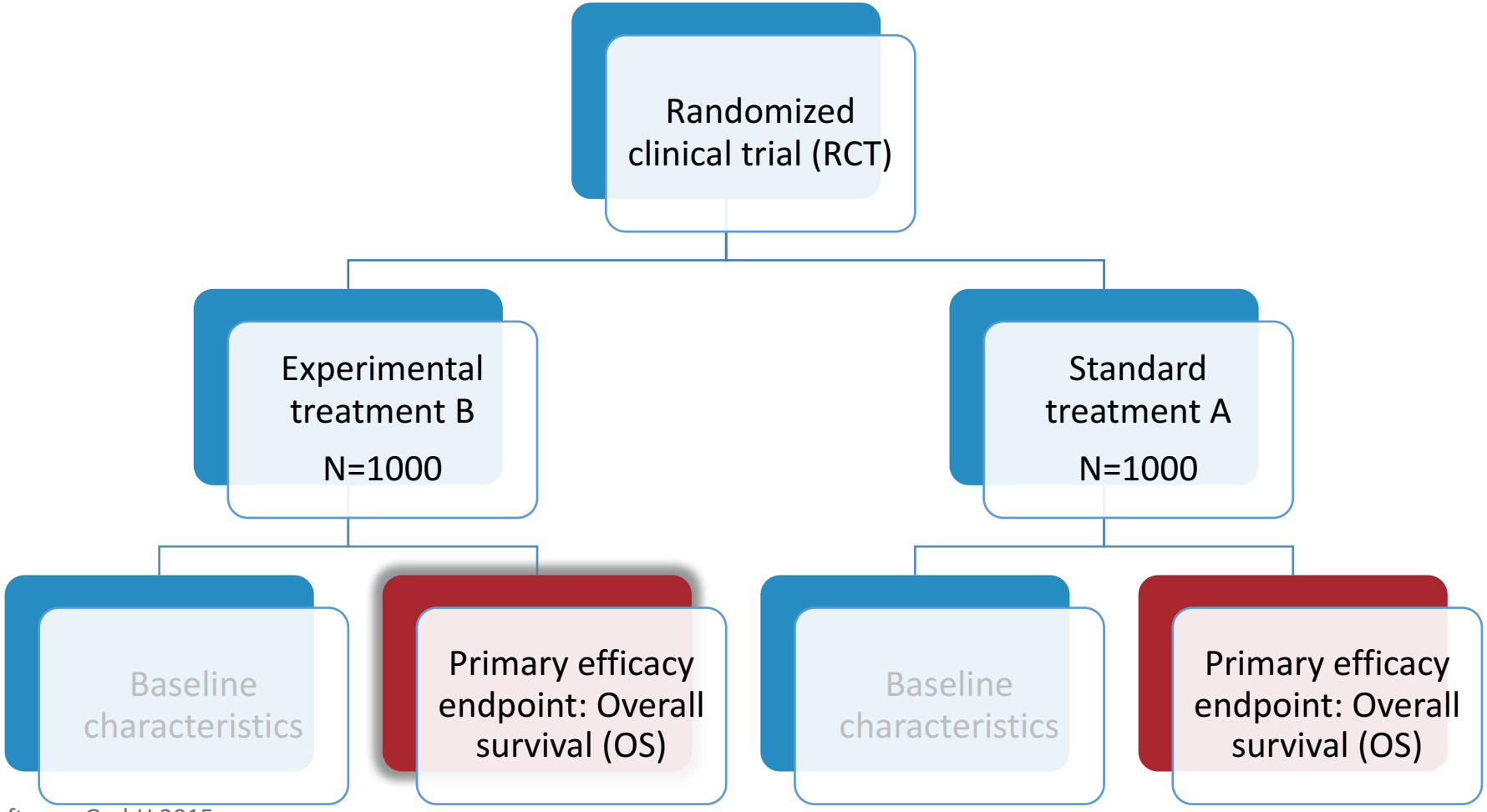
* from Cox proportional hazard regression model

- Significant interaction, only receptor-positive patients seem to benefit from the experimental treatment → however, limitations of subgroup analyses are still valid!

©HMS Analytical Software GmbH 2015

Next Simple Example

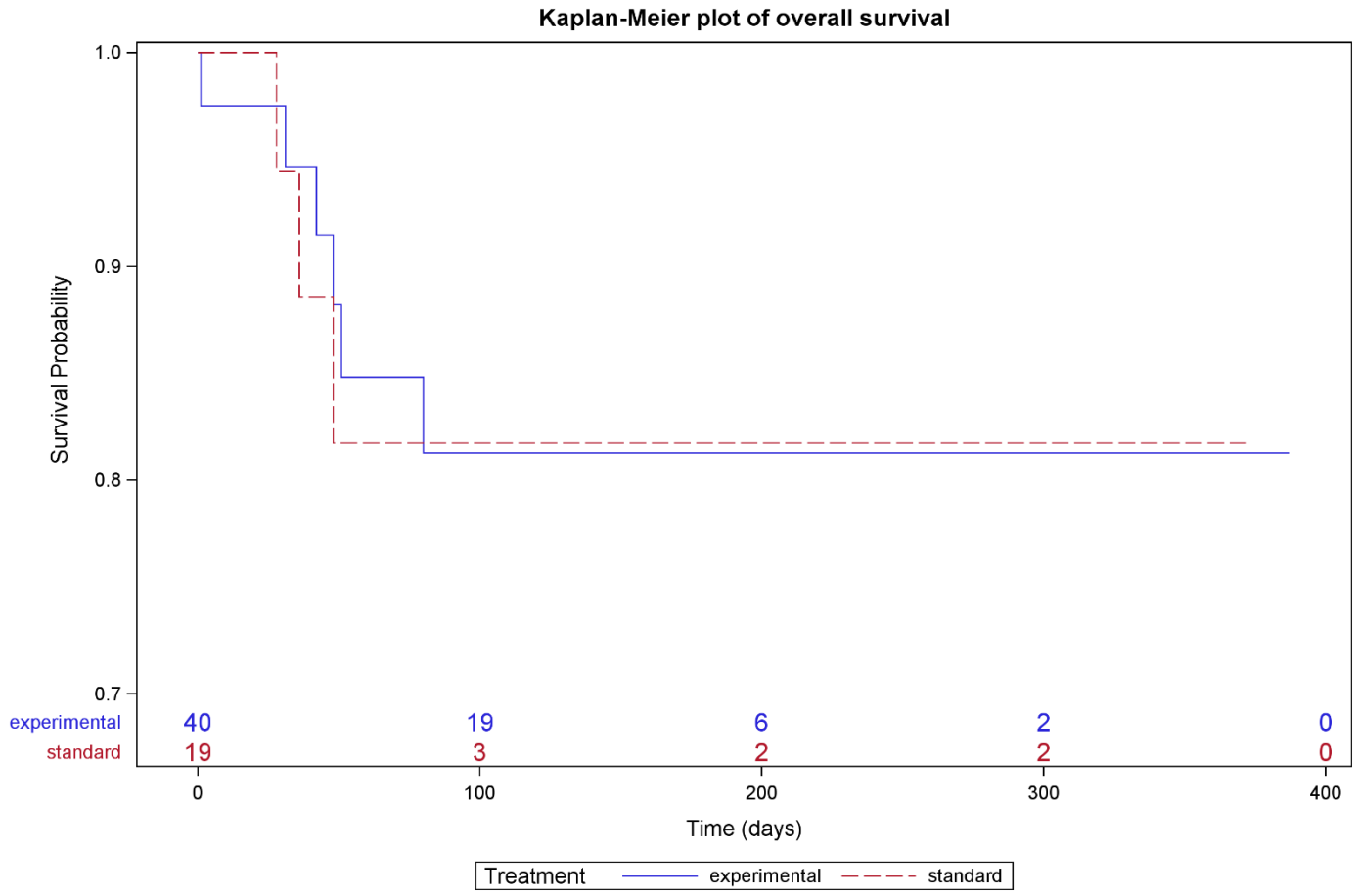
Primary endpoint



©HMS Analytical Software GmbH 2015

Next Simple Example

Kaplan-Meier Plot: Overemphasizing small difference

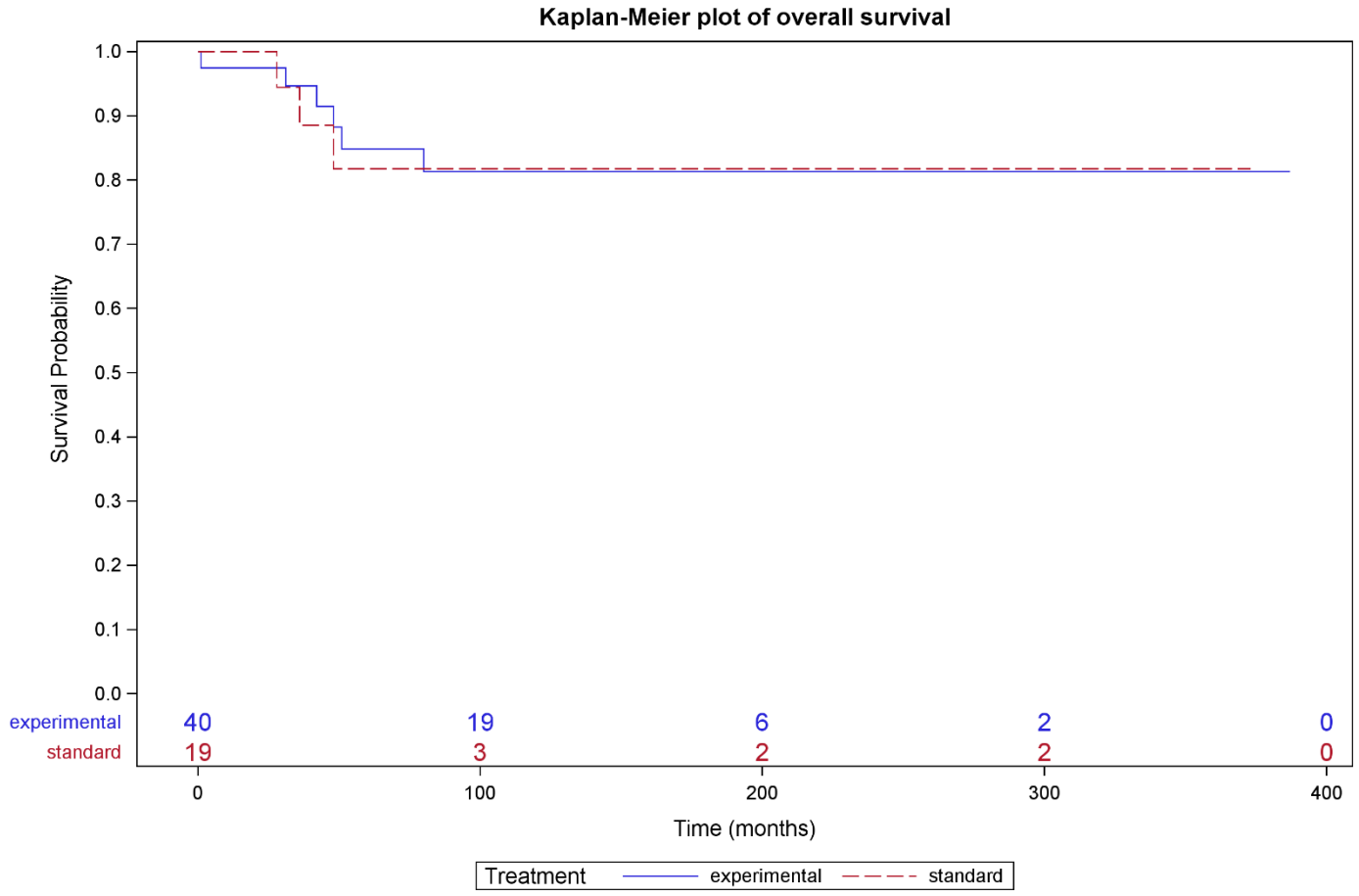


- y-axis only from 0.7 to 1.0
- Small difference appears more relevant than it actually is
- You cannot compare with other KM-plots

©HMS Analytical Software GmbH 2015

Next Simple Example

Kaplan-Meier Plot: Correct plot

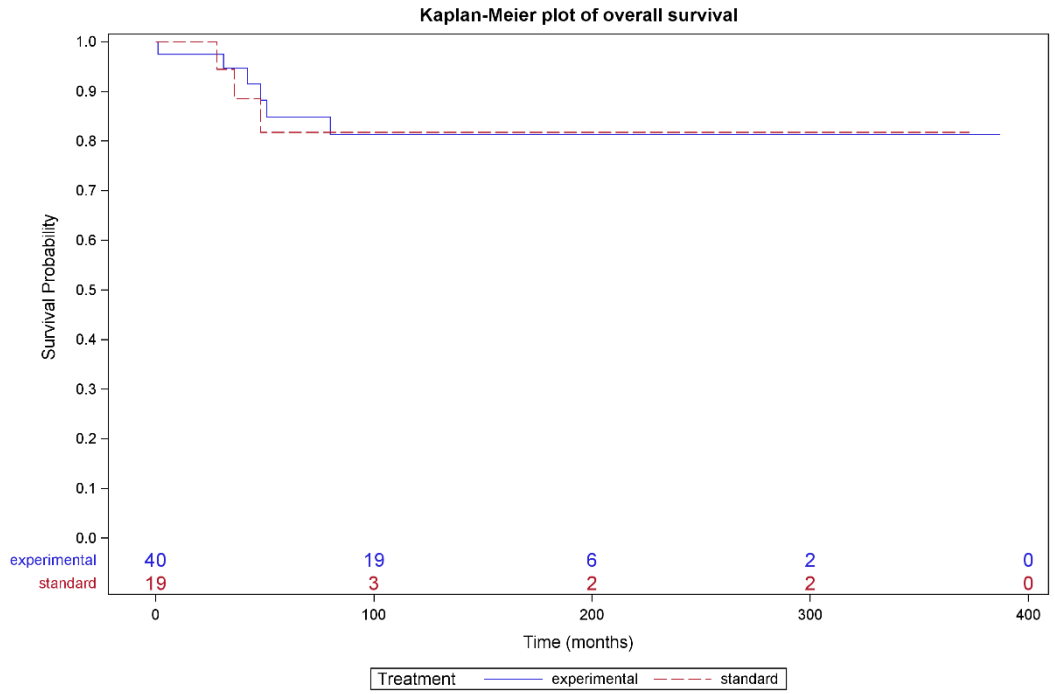


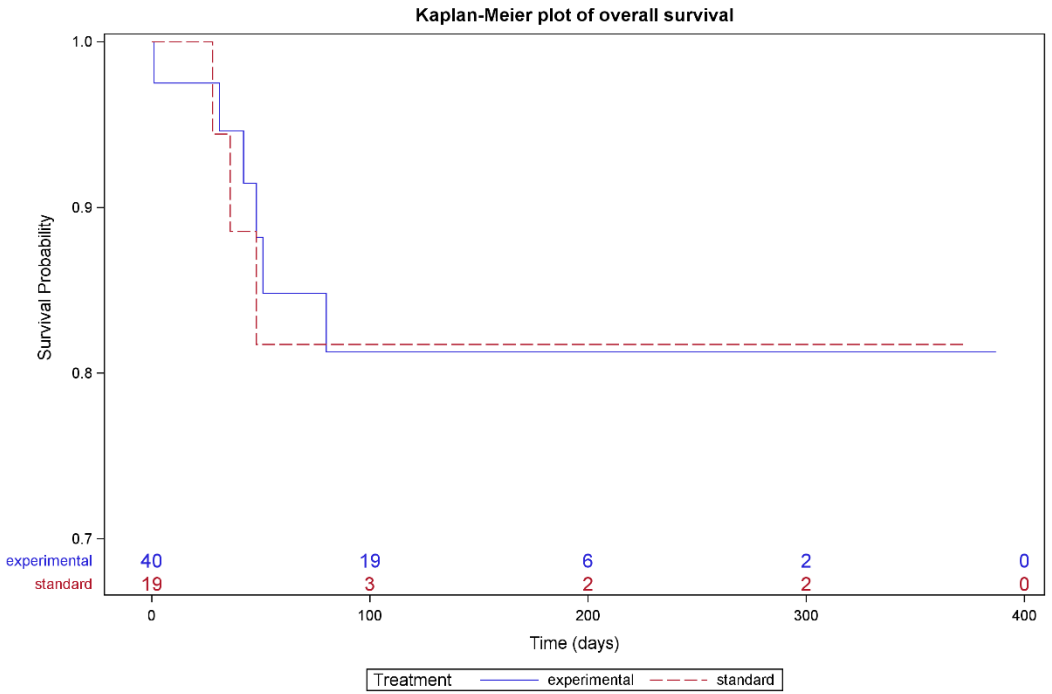
- y-axis should always go from 0 to 1
- Similarly, for all plots showing percentage/probability on the y-axis
- Sometimes hard to read if data lies only in a small range of the y-axis

©HMS Analytical Software GmbH 2015

Next Simple Example

Kaplan-Meier Plot: Good compromise





- Show both if it is important to show the small difference

©HMS Analytical Software GmbH 2015